HUMBOLDT-UNIVERSITÄT ZU BERLIN
INSTITUT FÜR BIBLIOTHEKS- UND INFORMATIONSWISSENSCHAFT

BERLINER HANDREICHUNGEN
ZUR BIBLIOTHEKS- UND
INFORMATIONSWISSENSCHAFT

HEFT 449

MULTILINGUALE HERAUSFORDERUNGEN IN DER SACHERSCHLIEßUNG

EIN RETRIEVALTEST AN EINEM MEHRSPRACHIGEN BIBLIOTHEKSBESTAND

VON
SARAH FALLERT

Multilinguale Herausforderungen in der Sacherschließung

Ein Retrievaltest an einem mehrsprachigen Bibliotheksbestand

von
Sarah Fallert

Berliner Handreichungen zur
Bibliotheks- und Informationswissenschaft

Begründet von Peter Zahn
Herausgegeben von
Vivien Petras
Humboldt-Universität zu Berlin

Heft 449

Fallert, Sarah

Multilinguale Herausforderungen in der Sacherschließung - ein Retrievaltest an einem mehrsprachigen Bibliotheksbestand / von Sarah Fallert. - Berlin : Institut für Bibliotheks- und Informationswissenschaft der Humboldt-Universität zu Berlin, 2020. - 107 S. : graph. Darst. - (Berliner Handreichungen zur Bibliotheks- und Informationswissenschaft ; 449)

ISSN 14 38-76 62

Abstract:

In einer Zeit, in der vermehrt automatische Verfahren der Inhaltserschließung eingesetzt und (weiter)entwickelt werden, leistet die vorliegende Studie einen Beitrag zur Diskussion um die Bedeutung der intellektuellen Sacherschließung beim Information Retrieval in einem multilingualen Kontext. Am Fallbeispiel des Online-Katalogs des Ibero-Amerikanischen Instituts wird in einem Retrievaltest die Beteiligung der überwiegend deutschsprachigen Schlagworte aus einem lokalen Thesaurus beim Auffinden von Dokumenten evaluiert. In der Studie werden 80 aus einem Logfile gewonnene Suchanfragen getestet, die in zwei gleich große Purpose Samples unterteilt wurden: ein deutsch- und ein fremdsprachiges.
Ein zentrales Ergebnis der Analyse ist, dass die Indexierung mit lokalen Schlagworten für einen erheblichen Anteil der zu den Suchanfragen aufgefundenen Dokumenten verantwortlich ist. Es lassen sich jedoch unter Berücksichtigung der Multilingualität der Suchanfragen deutliche Unterschiede mit Blick auf die Bedeutung der lokalen Schlagworte ausmachen. Auch das Verhältnis der intellektuellen Sacherschließung zur automatischen Indexierung von Elementen der Kataloganreicherung (Inhaltsverzeichnisse, Volltexte o.Ä.) wird näher beleuchtet und Stärken und Schwächen der verschiedenen Erschließungsformen werden diskutiert. Abschließend werden Möglichkeiten aufgezeigt, das Potential der lokalen Schlagworte insbesondere für fremdsprachige Suchanfragen stärker einzusetzen, um die zeit- und damit kostenintensive intellektuelle Sacherschließung sinnvoll nachzunutzen.

Diese Veröffentlichung geht zurück auf eine Masterarbeit im weiterbildenden Masterstudiengang im Fernstudium Bibliotheks- und Informationswissenschaft (Library and Information Science, M. A. (LIS)) an der Humboldt- Universität zu Berlin.

Eine Online-Version ist auf dem edoc Publikationsserver der Humboldt-Universität zu Berlin verfügbar.

Inhaltsverzeichnis

Abbildungsverzeichnis

Abkürzungsverzeichnis

AUTINDEX	Automatische Indexierung
BMBF	Bundesministerium für Bildung und Forschung
CACAO	Cross-Language Access to Catalogues and On-line Libraries
CBS	Zentrales Bibliothekssystem des GBV
CLEF	Conference and Labs of the Evaluation Forum; vorher: Cross-Language Evaluation Forum
CLR	Council on Library Resources
DFG	Deutsche Forschungsgemeinschaft
DNB	Deutsche Nationalbibliothek
GBV	Gemeinsamer Bibliotheksverbund
GESIS	Leibniz-Institut für Sozialwissenschaften
GND	Gemeinsame Normdatei
IAI	Ibero-Amerikanisches Institut Preußischer Kulturbesitz
IFLA	International Federation of Library Associations and Institutions
IR	Information Retrieval
IIR	Interactive Information Retrieval
KASCADE	Katalogerweiterung durch Scanning und automatische Dokumenterschließung
KI	Künstliche Intelligenz
KoMoHe	Kompetenzzentrum Modellbildung und Heterogenitätsbehandlung
LBS	Lokales Bibliothekssystem
LoC	Library of Congress (mit Sitz in Washington, D.C.)
LSW	Lokale Schlagworte
MILOS	Maschinelle Indexierung zur erweiterten Literaturerschließung in Online-Systemen
NEBIS	Netzwerk von Bibliotheken und Informationsstellen in der Schweiz
NLP	Natural Language Processing
OCLC	Online Computer Library Center

OCR	Optical Character Recognition; Optische Zeichenerkennung
OLC	Online Contents
OPAC	Online Public Access Catalog; Online-Katalog
OSIRIS	Osnabrücker Intelligent Research Information System
PETRUS	Prozessunterstützende Software für die digitale Deutsche Nationalbibliothek
PICA	Project for Integrated Catalogue Automation
PPN	PICA-Produktionsnummer; Identifikationsnummer für Datensätze des GBV
RSWK	Regeln für den Schlagwortkatalog
SPK	Stiftung Preußischer Kulturbesitz
SUB	Staats- und Universitätsbibliothek
SW	Schlagworte
SWD	Schlagwortnormdatei
TLA	Transaction-Logfile-Analyse
ToC	Table of Content; Inhaltsverzeichnis
TREC	Text Retrieval Conference
TU	Technische Universität
UB	Universitätsbibliothek
ULB	Universitäts- und Landesbibliothek
VZG	Verbundzentrale des GBV (mit Sitz in Göttingen)

Danksagung

Ich möchte mich von Herzen bei Britta Steinke, Benedikt Krüger, Freddy Álvarez sowie meinen Eltern bedanken, die mich bei der Bearbeitung meiner Masterarbeit und deren Überarbeitung zu dieser Publikation in vielfacher Hinsicht unterstützt und begleitet haben. Außerdem danke ich all den Kolleg_innen aus dem IAI, die mir stets mit ihren Fachkenntnissen zur Seite standen, insbesondere Ralf Ullrich, Rüdiger Stratmann, Francisca Roldán und Alexander Lozze. Ganz besonderer Dank gilt an dieser Stelle Magdalena Roos von der VZG, die immer sehr geduldig für meine vielen Fragen ansprechbar war und sich außerordentlich viel Zeit genommen hat, um diese zu beantworten.

1. Einleitung

> „La certidumbre de que algún anaquel en algún hexágono encerraba libros preciosos y de que esos libros preciosos eran inaccesibles, pareció casi intolerable.“[1]
>
> „Afirman los impíos que el disparate es normal en la Biblioteca y que lo razonable (y aun la humilde y pura coherencia) es una casi milagrosa excepción.”[2]

Diese beiden Zitate aus der Erzählung *La biblioteca de Babel* des argentinischen Schriftstellers Jorge Luis Borges, der zugleich einige Jahre Direktor der argentinischen Nationalbibliothek war, benennen zwei zentrale Aspekte, die auch die Frage nach Wert und Nutzen der Inhaltserschließung begleiten: zum einen das Ziel der (Wieder-)Auffindbarkeit von Medien innerhalb eines Bibliotheksbestands und zum anderen die kohärente und systematische Ordnung eben dieses Bestands, die jedoch von einigen „Gottlosen“ durchaus in Abrede gestellt wird.
Mit dem Titel der Erzählung wird zuletzt die dritte relevante Komponente der vorliegenden Arbeit erfasst. In der babylonischen Bibliothek, die Borges beschreibt, stehen die Sprache(n) und ihr Alphabet im Zentrum. Überträgt man die babylonische Sprachverwirrung auf die Recherche in Informationssystemen im 21. Jahrhundert, so lässt sie sich gut unter dem Begriff der Multilingualität fassen. Damit gemeint ist sowohl die Mehrsprachigkeit der Nutzer_innen als auch der Bibliotheksmedien.

Diesen drei Aspekten möchte die vorliegende Arbeit mit Blick auf den Einsatz von Schlagworten aus einem lokalen Thesaurus nachgehen. Es stellt sich damit die Frage nach der durch die intellektuelle Erschließung der Medien geschaffenen Ordnung, die deren Wiederauffindbarkeit zum Ziel hat, insbesondere auch dann, wenn unterschiedliche Sprachen bei der Formulierung der Suchanfragen zum Einsatz kommen.

---

[1] Borges 2005, S. 111. Die deutsche Übersetzung hierzu lautet: „Die Gewißheit, daß irgendein Regal in irgendeinem Sechseck kostbare Bücher berge, daß aber diese Bücher unzugänglich seien, erschien nahezu unerträglich“ (Borges 2007, S. 27).

[2] Borges 2005, S. 112. In der deutschen Übersetzung lesen wir: „Die Pietätlosen behaupten, daß in der Bibliothek der Unsinn an der Tagesordnung ist und daß das Vernunftgemäße (ja selbst das schlicht und recht Zusammenhängende) eine fast wundersame Ausnahme bildet“ (Borges 2007, S. 30).

Zwar verfolgt diese Arbeit einen systemorientierten Ansatz, der den Blick auf den Indexierungsprozess und die Retrievalergebnisse aus der Recherche im Online-Katalog legt; allerdings lassen sich aus den Ergebnissen einer quantitativ vorgehenden Evaluationsstudie durchaus auch Konsequenzen für Wert und Nutzen der lokalen Schlagworte für die Nutzer_innen ziehen.

Bezogen auf die drei eingangs benannten Aspekte zeigt sich ein ambivalentes Bild hinsichtlich der Bedeutung der lokal vergebenen Schlagworte, und gerade die Herausforderungen der Multilingualität zwingen zu einer differenzierten Sicht auf ihren Nutzen beim Retrieval. Das große Potential ihrer ordnenden Kraft bleibt jedoch bestehen und könnte sicherlich durch eine stärkere Berücksichtigung der Mehrsprachigkeit der Medien und Nutzer_innen noch weiter ausgeschöpft werden. Denn angesichts der „Sprachverwirrung“ in der babylonischen Bibliothek sollte nicht die blasphemisch unterstellte Unordnung an der Tagesordnung sein, sondern ein System, das über die Grenzen der Sprache(n) hinweg vermittelt und Brücken baut.

2. Theoretische Grundlagen

Bevor die Arten und Methoden der Sacherschließung sowie die damit verbundenen Aspekte der Multilingualität genauer beleuchtet werden, sollen im Folgenden einige der in dieser Arbeit gebrauchten Begriffe erläutert werden. Viele dieser Begriffe werden in der Forschungsliteratur nicht zwangsläufig einheitlich gebraucht, weshalb eine terminologische Klärung sinnvoll erscheint. Hinzu kommt, dass eine Vielzahl der konsultierten Literatur aus dem anglo-amerikanischen Raum stammt, der wiederum seine eigene, englischsprachige Terminologie anwendet, die sich nicht immer problemlos mit den deutschen Bezeichnungen korrelieren lässt.

Dokumente: Darunter verstanden werden nicht nur Texte jedweder Form (analog/digital, Monographien, Zeitschriften, Artikel usw.), sondern jedwedes Medium, das im Bibliotheksbestand enthalten ist, also auch Bild-, Audio- und Videomaterial. Synonym dazu wird vereinzelt auch von Medien gesprochen.

Multilingualität: Unter Multilingualität wird hier die Mehrsprachigkeit sowohl der Bibliotheksbestände als auch der Nutzer_innen verstanden: D.h. sowohl die Sprachen, in denen die Dokumente verfasst sind, als auch die von den Nutzer_innen bei der Recherche eingesetzten Sprachen variieren. Mehrsprachigkeit kann darüber hinaus auch bezogen auf jedes einzelne Dokument bzw. jede einzelne Person gegeben sein. So können Dokumente vorliegen, die Passagen in verschiedenen Sprachen enthalten, oder Nutzer_innen am Katalog recherchieren, die mehr als eine Sprache dominieren. Mehrsprachigkeit bildet damit das Gegenstück zur Einsprachigkeit (Monolingualität) (siehe dazu genauer Kapitel 2.2.).

Normsatz: Normsätze meinen hier die Normdaten, die in der Formalerschließung – z.B. zu Personen- und Körperschaftsnamen – und der Sacherschließung – in Form normierter Einträge zu den Elementen einer Dokumentationssprache – gleichermaßen hinterlegt werden. Sie können sowohl Fremddaten (z.B. der GND) als auch dem lokalen Thesaurus des IAI entstammen.

Sucheinstieg: Damit gemeint sind aus Nutzer_innensicht alle Zugangspunkte, die der Online-Katalog für die Recherche bietet, d.h. alle indexierten Katalogfelder sowie die Indexterme, die im Zuge der Kataloganreicherung aus den Volltexten extrahiert werden (zum Vorgang der Indexierung siehe Kapitel 2.1.). In dieser Arbeit werden 5 Sucheinstiege unterschieden, die gleichzeitig die Variablen der statistischen Auswertung bilden (siehe Kapitel 2.4.).

Suchterm: Unter Suchtermen werden die von den Nutzer_innen in die Suchmaske des Online-Katalogs eingegebenen Wörter verstanden, also die Bestandteile der Suchanfragen auf Wortebene. Synonym hierzu ist in der Forschungsliteratur von Stichworten (Keywords) die Rede. Je nach eingenommener Perspektive können Stichworte auch die im Zuge der Indexierung aus den Dokumenten und ihren bibliographischen Metadaten extrahierten Terme bezeichnen; in dieser Arbeit wird in diesem Fall von Indextermen oder Indexeinträgen gesprochen.

Schlagwort: Im Rahmen dieser Arbeit werden damit die aus kontrollierten Vokabularen vergebenen Deskriptoren bezeichnet (siehe dazu genauer Kapitel 2.1.). Dabei wird zwischen lokalen Schlagworten aus dem Thesaurus des IAI und solchen aus Fremddaten differenziert. Bei der Dokumentation der Tests wird dieser Begriff in einem etwas erweiterten Sinne angewandt, da bei den Schlagworten aus Fremddaten im Einzelfall auch Klassifikationen eine Rolle spielen können (z.B. die verbal verfassten Bezeichnungen der Haupt- und Unterklassen der Basisklassifikation, die ebenfalls indexiert werden). Beim Beschreibungsteil der Testdurchführung wird zudem die Abkürzung SW für „Schlagworte“ benutzt und für „lokale Schlagworte“ die Abkürzung LSW. Im Englischen ist die Bezeichnung Subject Heading üblich.

2.1. Arten und Methoden der Sacherschließung

Die inhaltliche Erschließung von Büchern und anderen Medien – auch Inhalts- oder Sacherschließung genannt – spielt seit jeher eine zentrale Rolle in Bibliotheken und anderen Informations- und Dokumentationseinrichtungen, sei es durch Findbücher, Aufstellungssystematiken, kontrollierte Vokabulare und Klassifikationen oder zunehmend durch automatisierte Verfahren wie die Volltextindexierung oder die Anreicherung durch eingescannte ToC (Table of Contents), Rezensionen usw.[3] Hintergrund für die Entstehung maschineller oder automatischer Verfahren[4] ist der Anstieg elektronischer Publikationen, der durch institutionelle oder disziplinäre Repositorien, die wachsende Open-Access-Bewegung oder die Retrodigitalisierung begünstigt wird und mit einer Auflösung der vormals klaren Grenzen zwischen dem Dokument selbst und seinem Katalogisat einhergeht.[5]

Diese verschiedenen Formen der Inhaltserschließung erfüllen teilweise unterschiedliche Zwecke und Funktionen, die im Zusammenhang mit den Bedürfnissen der Nutzer_innen zu sehen sind sowie mit den in einer Bibliothek verfügbaren Medien, die in Menge und Form stark variieren können.[6] Ein in seinem Umfang begrenzter Bibliotheksbestand – etwa die in Freihandaufstellung verfügbaren Medien einer kleinen Gemeindebibliothek – erfordert einen anderen Zugriff bei seiner thematischen Erschließung als der Bestand einer großen Universitätsbibliothek, der eine rapide wachsende Zahl elektronischer Ressourcen vorhält. Steht bei ersterem Beispiel die Systematisierung der Medien über ihre Aufstellung im Vordergrund, die zugleich das Browsing am Regal ermöglicht, so muss bei letzterem die thematische Systematisierung nachvollziehbar und kongruent im Online-Katalog oder dem eingesetzten Discovery-System erkennbar sein, etwa durch die Zuordnung von Schlagworten, Klassifikationen oder Tags und die Möglichkeit auch nachträglich innerhalb der Treffermengen zu filtern. Es stellt sich damit zuallererst immer die Frage nach dem zu erschließenden Bestand und Ziel und Zweck seiner (inhaltlichen) Erschließung sowie den Zielgruppen seiner Nutzung.

---

[3] Für einen historischen Abriss über die kooperative inhaltliche Erschließung in deutschen Bibliotheken siehe Siegmüller 2007, S. 46-48.

[4] Die Begriffe automatische und maschinelle Indexierung werden in diesem Kontext synonym verwendet, wobei ich mich hier an Siegmüller orientiere, die beide Begriffe dahingehend versteht, dass maschinenbasierte, automatisierte Vorgänge zur endgültigen Gewinnung der Indexterme führen; vgl. ebd., S. 26.

[5] Vgl. ebd., S. 56 f., 60-62.

[6] Bertram sieht als die drei zentralen Faktoren, die es bei der Erstellung eines Erschließungskonzepts zu beachten gilt: 1) die Analyse der Nutzer_innen, 2) die personellen, materiellen, finanziellen, zeitlichen und technischen Rahmenbedingungen, und 3) eventuelle Besonderheiten des Gegenstandsbereichs. Vgl. hierzu Bertram 2005, S. 28.

Ich möchte im Folgenden eine kurze Übersicht über verschiedene Formen der Inhaltserschließung geben, wobei v.a. die intellektuelle Erschließung mittels kontrollierter Vokabulare im Vordergrund stehen wird.

Bertram versteht unter Inhaltserschließung die „Gesamtheit der Methoden und Hilfsmittel zur inhaltlichen Beschreibung von Dokumenten“[7] und benennt als ihre Zielsetzungen die Wiederauffindbarkeit von Dokumenten, den schnellen Zugriff auf sie sowie eine Beschleunigung der Relevanzentscheidung.[8] Ein weiteres Hauptziel ist Bertram zufolge die Reduktion sprachlicher Mehrdeutigkeit, indem **Begriffe** und deren **Bezeichnungen** – d.h. Sachverhalte und ihre Ausdrucksformen – in eine klare Beziehung zueinander gesetzt werden (z.B. Synonymie oder Polysemie).[9] Weitere Herausforderungen für die Vereinheitlichung bzw. Systematisierung sprachlicher Ambiguität in der Inhaltserschließung ergeben sich durch Schreibweisenvielfalt (durch orthographische Veränderungen oder auch abweichende Schreibweisen zwischen verschiedenen Sprachen), Wortformenvielfalt, Ausdrucksvielfalt, Begriffskombinationen, implizite Inhalte, unklare Begriffsbeziehungen sowie die nur aus dem Kontext des Auftretens ersichtliche Relevanz einer Bezeichnung für den Dokumenteninhalt.[10] Bertram folgert daraus für die dokumentarische Tätigkeit folgende Aufgaben: Korrektur von Schreibfehlern, Normierung von Bezeichnungen, Zusammenführen von Synonymen, eindeutiges Aufschlüsseln von Homonymen und Polysemen, Offenlegen von Begriffsbeziehungen, Explizieren impliziter Inhalte, Überführen umgangssprachlicher oder metaphorischer Ausdrucksformen in formale Sprache, Lexikalisierung von Paraphrasen und Zerlegung von Begriffskompositionen sowie die Unterscheidung von Wichtigem und Unwichtigem.[11]

Grundsätzlich lässt sich die inhaltliche Erschließung in die Teilprozesse der **Inhaltsanalyse** und der **Inhaltsdarstellung** untergliedern[12] und auf verschiedenen Ebenen in unterschiedliche Methoden und Verfahren unterscheiden: Abstracting vs. Indexieren,

---

[7] Ebd., S. 18.

[8] Vgl. ebd., S. 18 f. Ähnliche Definitionen finden sich bei Siegmüller, laut der die inhaltliche Erschließung „[…] innerhalb eines definierten Bestandes Literatur zu einem bestimmten Thema zusammenführen soll“ (Siegmüller 2007, S. 7). An anderer Stelle finden wir außerdem folgende Definition: „Die inhaltliche Erschließung (auch Inhaltserschließung oder Sacherschließung genannt) dient dazu, den Inhalt von Dokumenten einer Sammlung so präzise zu beschreiben, dass sie bei einer thematischen Anfrage gezielt wieder gefunden werden können“ (ebd., S. 11).

[9] Vgl. Bertram 2005, S. 22 und ausführlicher Kapitel 2, S. 31-48. Begriffe stellen demnach gedankliche Abstraktionen dar, die in Äquivalenz-, Hierarchie- und Assoziationsbeziehungen eingebunden sind, während Bezeichnungen sprachliche Repräsentationen dieser abstrakten Sachverhalte sind, die sowohl natürlichsprachig sein können (z.B. Schlagworte oder Deskriptoren) als auch künstlichsprachig (etwa die Bezeichnung durch Nummern oder Notationen).

[10] Vgl. ebd., S. 42 f.

[11] Vgl. ebd., S. 46.

[12] Vgl. ebd., S. 22 f.

intellektuelle vs. automatische Inhaltserschließung, verbale vs. klassifikatorische Erschließung.[13]

Während beim **Abstracting** „die Inhalte von Dokumenten im Kontext wiedergegeben" werden, z.B. durch Anreicherung der Katalogisate mit kurzen Zusammenfassungen (Abstracts) oder mit Abbildungen der ToC, werden beim **Indexieren** „dem Dokument einzelne inhaltskennzeichnende Bezeichnungen zugeteilt".[14]

In dieser Arbeit liegt der Fokus auf der Erschließung durch Indexierung, die ihrerseits sowohl intellektuell als auch maschinell erfolgen kann. Bei der Indexierung kann grundsätzlich zwischen **Extraktions-** und **Additionsverfahren** unterschieden werden: Während bei der Extraktion die Indexterme dem (textbasierten) Dokument entnommen werden, können bei der Addition auch Terme zugewiesen werden, die nicht im Dokument selbst enthalten sind und einer Dokumentationssprache entstammen – z.B. in Form von Schlagworten.[15] Die dem Dokument zugewiesenen Bezeichnungen werden Indexterme genannt; das Ergebnis des Indexierungsvorgangs bildet das Indexat, das alle zu einem Dokument gehörenden Indexterme enthält.[16] Die den verschiedenen Dokumenten zugeordneten Indexterme werden in den Index (auf Deutsch auch als Register bezeichnet) aufgenommen und stellen als Repräsentationen der Inhalte der Dokumente die Grundlage für das Matching zwischen den Suchanfragen der Nutzer_innen und den in einer Sammlung enthaltenen Dokumenten dar. Während durch das Indexieren also Zugänge zu den Dokumenten geschaffen werden sollen, zielt das im Anschluss stattfindende **Information Retrieval** (IR) darauf ab, gezielt bestimmte Inhalte aufzufinden. Diese beiden Prozesse sind als komplementär und aufeinander bezogen zu verstehen und können mit Nohr folgendermaßen definiert werden: „Indexieren erfüllt den Zweck einer inhaltlichen Repräsentation von Dokumenten mit dem Ziel, diese im Zuge eines Information Retrieval unter entsprechenden Deskriptoren suchbar und auffindbar zu machen."[17]

Bei der **intellektuellen Erschließung** sind Dokumentationssprachen ein zentrales Instrument, um die zu beschreibenden Sachverhalte zu systematisieren, wobei zwischen

[13] Vgl. ebd., S. 24-26.

[14] Beide Zitate: ebd., S. 24.

[15] Vgl. hierzu ebd., S. 80 und Siegmüller 2007, S. 11. Unterschieden wird zudem zwischen Prä- und Postkoordination, also der Entscheidung die begrifflichen Komponenten eines Sachverhalts entweder bereits gemeinsam zu indexieren und auch nur in Kombination suchbar zu machen (Präkoordination) oder voneinander getrennt zu indexieren und erst im Retrieval wieder zusammenzuführen (Postkoordination). Vgl. Bertram 2005, S. 74 f. und Siegmüller 2007, S. 44 f.

[16] Vgl. Bertram 2005, S. 67 f.

[17] Nohr 2001, S. 15. Bertram unterscheidet einen weiten und einen engen Begriff des Information Retrieval: einerseits als „die Repräsentation, Speicherung und Organisation von sowie den Zugriff auf Informationen, die in Retrievalsystemen gespeichert werden können" und andererseits als „das methodisch geleitete, gezielte Wiederauffinden von Dokumenten" (beide Zitate: Bertram 2005, S. 19).

einer gröberen Ordnung durch **Klassifikationen** und einer feineren Ordnung durch **Thesauri** und die in ihnen enthaltenen Deskriptoren unterschieden werden kann.[18] Dokumentationssprachen, auch als kontrollierte Vokabulare bezeichnet, können mit Bertram definiert werden als: „die Gesamtheit aller Begriffe und ihrer sprachlichen Ausdrücke, die, nach bestimmten Regeln angewandt, vor allem dem Indexieren dokumentarischer Bezugseinheiten und ihrer gezielten Wiederauffindung dienen.“[19] Thesauri, als eine Ausprägung kontrollierter Vokabulare, zeichnen sich demgegenüber durch eine inhaltliche Feinerschließung sowie eine „geordnete Zusammenstellung von Begriffen und Benennungen, die zum Indexieren, Speichern und Wiederauffinden dokumentarischer Bezugseinheiten dient“,[20] aus und arbeiten mit den Prinzipien von Normung (Vereinheitlichung der Schreibweise und Ansetzung), terminologischer Kontrolle (Disambiguierung von Polysemie und Homonymie, Herstellung von Äquivalenzrelationen bei Synonymen, Entscheidung zur Zerlegung von Komposita) und begrifflicher Kontrolle (Festlegung von hierarchischen und assoziativen Beziehungen zwischen Begriffen).[21] Anders als systematische **Notationen**, die mit Buchstaben, Symbolen oder Ziffern arbeiten, sind **Deskriptoren** verbalsprachlich. **Schlagworte**, die bei ihrer Einbindung in einen Thesaurus zu Deskriptoren werden, stellen einen eindeutigen Repräsentanten eines Begriffes dar, der kurz und prägnant, dabei jedoch möglichst präzise und vollständig beschrieben wird und im Dokument vorkommen kann aber nicht muss.[22]

Thesauri sind jedoch nicht nur aufwändig in ihrem Aufbau, sondern insbesondere auch in ihrer Pflege, die aufgrund möglicher Veränderungen auf Begriffs- und Bezeichnungsebene erforderlich wird und kontinuierlich stattfinden muss; wird ein Deskriptor verändert oder entfernt, so wirkt sich dies auf das gesamte dokumentationssprachliche Beschreibungsnetz aus.[23]

---

[18] Vgl. ebd., S. 27.

[19] Ebd., S. 127.

[20] Ebd., S. 209.

[21] Vgl. ebd., S. 218-226 und Siegmüller 2007, S. 12.

[22] Vgl. Bertram 2005, S. 68 und Siegmüller 2007, S. 12. Der Begriff des Deskriptors kann unterschiedlich eng oder weit gefasst werden. So können diese ausschließlich als Beschreibungselemente von Thesauri verstanden werden oder alle natürlichsprachigen Indexterme bezeichnen, sodass in letzterem Fall nur Notationen aus ihrer Definition ausgenommen werden; vgl. Bertram 2005, S. 68. Im RSWK werden Schlagwort, Deskriptor, Vorzugsbezeichnung und Normierter Sucheinstieg synonym verwendet; vgl. Scheven/Nadj-Guttandin 2017, S. 36. Von den gebundenen Indexierungsverfahren, also solchen, die auf kontrollierte Vokabulare zurückgreifen, ist das freie Indexieren zu unterscheiden, bei dem freie, nicht normierte Vokabulare zugewiesen werden; vgl. Bertram 2005, S. 81 f.

[23] Vgl. Bertram 2005, S. 228-230. Bertram sieht die niedrige Abstraktionsstufe von Thesauri als Grund für ihre besondere Anfälligkeit angesichts der Entwicklung von Sprachen und Begriffsbedeutungen.

Mit dem Aufkommen des Semantic Web spielen zudem auch **Ontologien** als einer weiteren Form kontrollierter Vokabulare zunehmend eine Rolle bei der Inhaltserschließung von Dokumenten.[24]

Die **automatischen Indexierungsverfahren** hingegen arbeiten ausgehend von der Sprachoberfläche mit einer Extraktion von Stichworten (Indextermen) aus den digital(isiert) vorliegenden Dokumenten oder Dokumentteilen. Sie lassen sich differenzieren in: 1) statistische Verfahren, 2) computerlinguistische Verfahren, 3) begriffsorientierte Verfahren.[25] Bei statistischen Verfahren wird die Relevanz eines Indexterms ausgehend von der Häufigkeit seines Auftretens berechnet; wird das Verhältnis von dokumentspezifischer Termfrequenz und Dokumentfrequenz berücksichtigt, lässt es sich in Form der inversen Dokumenthäufigkeit ausdrücken. Computerlinguistische Verfahren arbeiten darüber hinaus mit einer weiteren Bearbeitung der Terme, etwa durch Tokenisierung, Spracherkennung, Stoppworteliminierung, Erkennung von Schreibfehlern und abweichenden Schreibweisen oder Wortformenreduktion (Stemming); sie können dabei sowohl wörterbuchbasiert als auch regelbasiert vorgehen. Begriffsorientierte Verfahren versuchen demgegenüber eine Annäherung an die Ebene der intellektuellen Erschließung, indem den extrahierten Termen mittels Erkennungswörterbüchern Deskriptoren zugewiesen und die Terme durch eine Auswertung der in ihrem Umfeld vorkommenden Wörter in ihrer Bedeutung disambiguiert werden.

---

[24] Vgl. Ingwersen/Järvelin 2005, S. 132. Zu diesem Thema siehe z.B. die Arbeit von Boltzendahl 2003.

[25] Zu diesen drei Verfahren vgl. Kempf 2013, S. 98 f. und Siegmüller 2007, S. 27-38. Eine gute Einführung in die automatische Indexierung bietet Nohr 2001, der neben den genannten Verfahren auch das Pattern Matching, also die Mustererkennung, als ein weiteres automatisches Verfahren aufführt; vgl. Nohr 2001, S. 23 f.

Wichtige Projekte im Bereich der automatischen Indexierung entstanden in Deutschland bereits seit den 1980er Jahren, beispielsweise das begriffsorientiert arbeitende Projekt AIR/PHYS, das zwischen 1981-1986 an der Technischen Hochschule Darmstadt durchgeführt wurde; vgl. Siegmüller 2007, S. 37. Zentral in diesem Zusammenhang sind die maßgeblich von Klaus Lepsky seit Mitte der 1990er Jahre an der ULB Düsseldorf durchgeführten DFG-geförderten Projekte MILOS I, MILOS II und KASCADE, die zu dem Ergebnis kamen, dass die automatische Indexierung unter Zuhilfenahme computerlinguistischer Bearbeitungsschritte (MILOS I), der Einbindung in die SWD (MILOS II) und zuletzt der statistisch basierten Termgewichtung (KASCADE) eine Erhöhung des Recalls relevanter Treffer bei geringen Einbußen an Precision und einer erheblichen Reduktion von Nulltreffern mit sich bringt; vgl. Oberhauser/Labner 2003, S. 306; Siegmüller 2007, S. 68-73 sowie eingehender zum Projekt MILOS z.B. Lepsky 1994 und zu KASCADE Lepsky/Zimmermann 2000. Siegmüller kommt zu dem Fazit, dass beide Projekte insgesamt positive Ergebnisse hervorbrachten, in der Folge jedoch nur wenig in Bibliotheken eingesetzt wurden, was unter anderem darauf zurückzuführen sei, dass die eingesetzten Bibliothekssysteme noch nicht in der Lage dazu gewesen seien, die volle Indexierungsleistung der Programme auch umzusetzen; vgl. Siegmüller 2007, S. 8, 67, 73. Ähnlich arbeitete das ebenfalls DFG-geförderte Projekt OSIRIS, das in den 1990er Jahren u.a. an der SUB Bremen und im schweizerischen Bibliotheksnetzwerk NEBIS durchgeführt wurde und mit einem natürlichsprachigen Ansatz, computerlinguistischen Komponenten sowie automatisch erweiterbaren Lexika den Recall für Suchanfragen in OPACs steigern konnte und möglichen Unschärfen in den thematischen Suchanfragen durch die Anzeige verschiedener passender Notationen entgegenwirkte; vgl. hierzu ebd., S. 75-84 sowie Recker/Ronthaler/Zillmann 1996 und Ronthaler/Zillmann 1998.

Einen Mittelweg bilden **computerunterstützte Erschließungsverfahren**, bei denen maschinelle Vorgänge, etwa maschinell generierte Vorschläge, lediglich Vorleistungen bilden, die im Anschluss intellektuell geprüft werden.[26]

Die Zielsetzungen der intellektuellen gegenüber der automatischen Sacherschließung lassen sich in Anlehnung an Nohr und Kempf folgendermaßen gegenüberstellen: Während die intellektuelle Sacherschließung eine korrekte und konsistente Repräsentation des zu erschließenden Mediums anstrebt, die systematisch oder verbal sein kann, steht bei der automatischen Inhaltserschließung die Wiederauffindbarkeit des Mediums im Vordergrund.[27] Die intellektuelle Sacherschließung erreicht die Bedeutungsebene der Dokumente, die in einem zweistufigen Prozess zunächst inhaltlich analysiert und dann in eine Dokumentationssprache übertragen werden. Demgegenüber agiert die automatische Inhaltserschließung „benennungsorientiert“[28] auf der sprachlichen Oberfläche.
Für die Auffindbarkeit der erschlossenen Medien sind in einem zweiten Schritt die jeweils zu Grunde gelegten **Retrievalmodelle** relevant, die je nach angewandter Erschließungsform gewählt werden. Bei der intellektuellen Erschließung werden zumeist das Exact-Match-Modell und Boole´sches Retrieval gewählt gegenüber Best- oder Partial-Match-Modellen verbunden mit Relevance Ranking beim Einsatz automatischer Verfahren.[29]

Die Vorteile der intellektuellen Erschließung liegen in ihrer „begriffsorientierten“[30] Vorgehensweise begründet, die von der Zeichenebene abstrahiert und zum Ziel hat, die Essenz eines Dokuments zu formulieren. Die in der inhaltlichen Erschließung eingesetzten kontrollierten Vokabulare leisten außerdem eine Herstellung semantischer Beziehungen, die Gleiches zusammenführt und Ungleiches voneinander unterscheidet, etwa durch die Benennung von Synonymen, Hierarchien und Assoziationen sowie die Bedeutungsdifferenzierung von Polysemen und Gebrauchskontexten. Die intellektuelle Inhaltserschließung ist außerdem auf alle Medienarten anwendbar, d.h. auch auf nicht textbasierte Medien wie Bild- oder Tonmaterial.[31]

---

[26] Vgl. Bertram 2005, S. 83 und Siegmüller 2007, S. 26 f. Ein Beispiel hierfür wäre der „digitale Assistent“ und sein Nachfolger DA-2, der seit 2013 an der Zentralbibliothek Zürich eingesetzt wird; vgl. dazu u.a. Hinrichs et al. 2016 und Malits/Schäuble 2014.

[27] Vgl. Nohr 2001, S. 16 und Kempf 2013, S. 98.

[28] Bertram 2005, S. 83.

[29] Vgl. Nohr 2001, S. 87-95; Siegmüller 2007, S. 15-20; Bertram 2005, S. 84 sowie Kempf 2013, S. 98. IR-Systeme, die auf Suchmaschinentechnologie aufbauen, verwenden algebraische (z.B. das Vektorraum-Modell) oder probabilistische Modelle.

[30] Bertram 2005, S. 83.

[31] Vgl. ebd., S. 83.

Als ein Nachteil intellektueller Erschließungsformen kann gesehen werden, dass dieses Verfahren sehr zeit- und personalintensiv ist, was aus Sicht der Bibliotheken einen zentralen Kostenfaktor bedeutet. Auch der Aspekt der Indexierungskonsistenz[32] kann sich als problematisch erweisen, da die intellektuelle Zuweisung beispielsweise von Schlagworten immer personengebunden erfolgt und damit der persönlichen Einschätzung des/der Indexierenden unterliegt. Ein gewisses Maß an Subjektivität bei der Zuordnung ist damit unausweichlich. Dem normierenden und vereinheitlichenden Charakter der eingesetzten kontrollierten Vokabulare, die sprachliche Ambiguität verringern sollen, steht so die Ambiguität der individuellen Ermessungsentscheidung einer einzelnen Person entgegen.

Für das **Retrieval** ist die Sacherschließung insofern relevant als sie über die Metadaten aus der Formalerschließung hinaus zusätzliche Sucheinstiege bietet, die zudem gezielt thematisch orientiert sind. Auch wenn Schlagworte selten von den Nutzer_innen direkt als Sucheinstieg genutzt werden[33] – etwa durch eine gezielte Schlagwortsuche, das Browsing in Schlagwortlisten usw. –, so liefern sie doch Indexterme, die auch in der thematischen Suche mit Stichworten aufgefunden werden können, selbst wenn dieser Vorgang den Nutzer_innen verborgen bleibt und ihnen häufig nicht bewusst ist.[34]

Ein zentraler Aspekt, den es jedoch beim Einsatz kontrollierter Vokabulare zu beachten gilt, ist, dass diese einer kontinuierlichen, sehr zeitaufwändigen Pflege bedürfen, um das eingesetzte Vokabular aktuell zu halten. Denn nur so können Deskriptoren Bedeutungsunterschiede und –veränderungen im Sprachgebrauch abbilden und tatsächlich einen Mehrwert im Vergleich zu einer auf der Sprachoberfläche agierenden Suche nach einer bestimmten Zeichenfolge generieren.[35]

Auch die automatische Indexierung schafft neue Sucheinstiege, häufig in großer Menge, z.B. durch die Extraktion von Termen aus Volltexten oder ToC. Diese Sucheinstiege sind anders als kontrollierte Vokabulare jedoch nicht normiert. Eine weitergehende Bearbeitung, etwa durch die oben erwähnten computerlinguistischen Verfahren ist jedoch möglich.

Eine Folge der Automatisierung ist, dass die inhaltliche Erschließung kleinteiliger ausfällt, da nun auch Ebenen eines Dokuments erschlossen werden, die bei der intellektuellen Erschließung zu Gunsten der thematischen Gesamtübersicht vernachlässigt werden. Während Schlagworte oder Klassifikationen ein Werk als Ganzes erfassen wollen und das

[32] U.a. Garrett hebt auf den Aspekt der Konsistenz bei der Vergabe von Schlagworten ab, die auch maßgeblich von der Pflege und Aktualität des kontrollierten Vokabulars abhinge; vgl. Garrett 2007, S. 69 f.

[33] Zu diesem Aspekt siehe Kapitel 2.6., in dem auf die Forschung zur Nutzung der Schlagwortsuche durch die Nutzer_innen genauer eingegangen wird.

[34] Garrett sieht gerade hierin das große Potential von kontrollierten Vokabularen, da sie die Stichwortsuche um sehr nützliche, häufig nur in den Schlagworten indexierte Terme erweitere; vgl. ebd., S. 71 f.

[35] Vgl. ebd., S. 70.

einschlägigste Thema festlegen, ermöglicht die Indexierung z.B. von ToC eine Erschließung auf der Ebene einzelner Kapitel oder Aufsätze. Sehr spezifische, für das Werk als Gesamtes weniger relevante Aspekte können so im Retrieval aufgefunden werden, was in der Folge zu einer Steigerung der Nutzung des Gesamtbestands führen kann.[36]

Nicht zuletzt die in jüngerer Zeit wieder verstärkt geführte Debatte um das Inhaltserschließungskonzept der DNB zeigt aber auch, dass die Automatisierung mit ihren Möglichkeiten, große Mengen an Dokumenten in wenig Zeit und auf kleineren Ebenen zu erschließen, zugleich an gewisse Grenzen stößt bzw. Begleiterscheinungen mit sich bringt, die je nach eingenommener Perspektive nicht wünschenswert erscheinen – insbesondere dann, wenn mit der Automatisierung ein Abschied von den traditionellen, intellektuellen Formen der Sacherschließung verbunden ist.[37] V.a. der 2006 eingeführte, erweiterte Sammelauftrag der DNB für Netzpublikationen, deren intellektuelle Erschließung 2010 eingestellt wurde, führte in der DNB zu einer flächendeckenden Einführung der maschinellen Inhaltserschließung, teilweise auch analoger Publikationsformen; wegweisend war hier das 2013 gestartete Projekt PETRUS, dessen zentrale Punkte die maschinelle Sachgruppen- sowie Schlagwortvergabe und die Weiternutzung von Inhaltserschließungsdaten waren.[38]
Eine Übersicht über den Stand der (semi-)automatisierten Klassifizierung im deutschsprachigen Raum bietet Kasprzik. Die Autorin kommt zu dem Schluss, dass die Anwendung solcher (semi)automatisierten Verfahren bei andernfalls gar nicht erschlossenen Beständen in jedem Fall sinnvoll sei, regt ansonsten jedoch an, Kosten und Nutzen insbesondere lernender maschineller Verfahren gut abzuwägen, da das Training der Maschinen sehr aufwändig sei und – ebenso wie die Kontrolle der Ergebnisse – menschliche

[36] Vgl. Siegmüller 2007, S. 57.

[37] Vgl. dazu etwa die Kritik von Ceynowa 2017, o.S. Klaus Ceynowa argumentiert darin u.a., dass gerade angesichts der viralen Verbreitung von Erschließungsdaten durch die stark gestiegene Übernahme von Fremddaten oder die Weiterverbreitung im Netz die Qualität dieser Metadaten umso zentraler werde. Zu der Aktualität der Debatte siehe auch das 2018 erschienene Themenheft der Zeitschrift BuB mit dem Titel *Ordnung schaffen: Die Zukunft der Sacherschließung* (Jahrgang 70, Heft 1). Rädler hingegen sieht die Indexerweiterung in Zeiten eines wachsenden Anteils elektronischer Ressourcen als unumgänglich und ein ausschließliches Festhalten an der intellektuellen Sacherschließung auch unter betriebswirtschaftlichen Aspekten als Fehlinvestition, da man es dergestalt verpasse, vorhandene Informationen wie etwa die in den ToC enthaltenen Terme zugänglich zu machen; vgl. Rädler 2004, S. 927, 938. Ebenfalls interessant in diesem Kontext ist der Beitrag von Alice Keller, der die Einstellungen zur (automatischen) Sacherschließung im deutsch- und englischsprachigen Raum untersucht; vgl. Keller 2015. In der von ihr durchgeführten Umfrage kommt die Autorin zu dem Ergebnis, dass in beiden Sprachräumen die Sacherschließung weiterhin ein hohes Ansehen genießt und aus Sicht der Befragten in der bisherigen Qualität aufrecht erhalten werden sollte, während automatisierte Verfahren eher skeptisch betrachtet und mit einem Qualitätsverlust gleichgesetzt werden; vgl. ebd., S. 811-813. Wyly beschreibt die Entscheidung zwischen einem Festhalten an der intellektuellen Sacherschließung und dem Verzicht auf diese als eine Wahl zwischen dem Herabsetzen von Katalogisierungsstandards gegenüber der Bildung von Rückstaus aufgrund der Zeitintensität dieses Verfahrens; vgl. Wyly 1996, S. 212.

[38] Vgl. hierzu u.a. Junger 2015; Schöning-Walter 2010 und 2011; Uhlmann 2013 sowie Wiesenmüller/Hinrichs 2017. Die automatische Zuweisung kontrollierter Vokabulare in der digitalen Bibliothek Europeana untersuchen Stiller et al. 2014.

Arbeit erfordere.[39] Sie merkt außerdem an, dass die sprachliche Homogenität der Trainingsdaten beim maschinellen Lernen die Ergebnisse deutlich verbessern könne, etwa durch Nutzung von Normdaten,[40] d.h. in der Regel wiederum menschlich erzeugte Vereinheitlichungen und Systematisierungen zusammengehöriger Informationen zu einer Person/einem Ort/einer Körperschaft/einem Schlagwort usw.

Siegmüller fasst die Vorteile und Schwächen automatisierter Verfahren folgendermaßen zusammen: Automatische Verfahren leisten eine schnelle und konsistente Bearbeitung großer Datenmengen und können das Retrieval durch linguistische Bearbeitung und Vereinheitlichung verbessern sowie mit Hilfe statistischer Verfahren das exakte Boole´sche Retrieval und die damit verbundenen Anwendungsprobleme überwinden. Nutzbringend sind sie v.a. dort, wo die personellen Ressourcen für die intellektuelle Inhaltserschließung fehlen sowie bei größeren Rückständen noch nicht erschlossener Dokumente. Probleme können durch das Agieren an der Sprachoberfläche entstehen, da dieses nicht auf dem Verstehen der Texte basiert und in der Folge v.a. die gewichtenden Algorithmen und die Aktualität der eingesetzten Wörterbücher und Lexika für die Güte der Retrievalergebnisse entscheidend sind. Zudem erfordern automatische Verfahren einen hohen technischen Aufwand und Input an Wissen. Auch Siegmüller plädiert darum für ein Nebeneinander beider Verfahren, da die Ergebnisse dort besonders gut ausfielen, wo automatisierte Verfahren auf intellektuell erschlossenem Material aufbauten.[41]

[39] Vgl. Kasprzik 2014, S. 103-106.

[40] Vgl. ebd., S. 104.

[41] Zu diesem Abschnitt vgl. Siegmüller 2007, S. 38 f.

2.2. Suchanfragen in multilingualen Kontexten

Mit dem Aufkommen des WWW und der Digitalisierung stehen den Nutzer_innen heute weltweit Dokumente in vielen verschiedenen Sprachen unmittelbar zur Verfügung. Die Frage der sprachlichen Zugänglichkeit stellt sich damit weit ausgeprägter und erfordert allein durch die schiere Menge verfügbarer Daten neue Ansätze. Die Forschung hat hier bereits einige Schritte unternommen, doch noch bestehen viele technische Grenzen bzw. Mängel oder Desiderate, die einer weiteren Auseinandersetzung bedürfen.[42]

Multilingualität stellt für das Retrieval eine besondere Herausforderung dar und begegnet uns bei allen Schritten des Rechercheprozesses. Ganz grundsätzlich kann man zwischen der Mehrsprachigkeit der Bibliotheksbestände und ihrer Metadaten aus der Erschließung auf der einen Seite und der Mehrsprachigkeit der Nutzer_innen auf der anderen unterscheiden.
Stiller/Gäde/Petras 2013 unterscheiden am Beispiel der digitalen Bibliothek Europeana folgende Ebenen der Multilingualität, die sowohl die Dokumente und das System als auch die Nutzer_innen betreffen: 1) Display, 2) Suche und Browsing, 3) Darstellung und Übersetzung der Ergebnisse, 3) Interaktion mit den Nutzer_innen.[43] Auf der Seite der Dokumente und des Systems lässt sich darüber hinaus zwischen multilingualen Interfaces, multilingualen Objekten und multilingualen Metadaten differenzieren.[44] Auf den Ebenen der Suche und der Repräsentation der Ergebnisse spielen besonders die Sprachen der Suchanfragen, die Indexierungssprachen – z.B. bei Nutzung multilingualer Thesauri – und die Sprachen, in der die Ergebnisse ausgegeben werden, eine Rolle.[45]
Die Mehrsprachigkeit der Nutzer_innen kann sich beim Retrieval ebenfalls auf verschiedenen Ebenen bemerkbar machen, begonnen bei der Auswahl einer Sprachvariante des Interfaces über die Einschränkung der Dokumente auf bestimmte Sprachen durch entsprechende Filter bis zur Eingabesprache der Suchanfragen und, nach erfolgreicher Suche, die Sprache der durch die Nutzer_innen ausgewählten Dokumente.[46]

---

[42] Vgl. dazu Oard 2009, o.S. Der Autor gibt zudem einen kurzen historischen Abriss über die Entwicklungen im Multilingual Information Access (MLIA), wobei er insbesondere die mit dem Ende des Kalten Krieges einsetzende Internationalisierung, das Aufkommen von WWW und Suchmaschinentechnologie sowie technische Entwicklungen wie z.B. maschinelle Übersetzungen hervorhebt. Laut Ingwersen/Järvelin setzt eine stärkere Auseinandersetzung mit Cross-Language Information Retrieval (CLIR) in den 1990ern ein; vgl. Ingwersen/Järvelin 2005, S. 121.

[43] Vgl. Stiller/Gäde/Petras 2013, S. 87 sowie ausführlicher zu weiteren Aspekten dieser Ebenen Gäde/Petras/Stiller 2010, S. 73-75.

[44] Vgl. Stiller/Gäde/Petras 2013, S. 87. Im Hinblick auf die Multilingualität der Metadaten beim Fallbeispiel Europeana siehe auch Stiller/Király 2017.

[45] Vgl. Gäde/Petras/Stiller 2010, S. 73-75.

[46] Vgl. ebd.

Zu unterscheiden ist des Weiteren zwischen multilingualem IR (MLIR) und cross-lingualem IR (CLIR). Der Unterschied zwischen diesen beiden Formen des IR liegt darin begründet, welcher Bereich von Mehrsprachigkeit betroffen ist: Im Fall von MLIR können sowohl die Anfragen als auch die Suchergebnisse in verschiedenen Sprachen vorliegen, während bei CLIR die Anfragen in einer Sprache erfolgen und nur die Ergebnisse multilingual sind.[47] Auch im Hinblick auf die jeweilige Zielgruppe lässt sich genauer differenzieren, auf welcher Ebene die Multilingualität angesiedelt ist. So lassen sich die Nutzer_innen nach Oard in „monoglots" und „polyglots" differenzieren, d.h. in Personen, die nur eine Sprache dominieren gegenüber solchen, die mehrere Sprachen sprechen.[48] Mehrsprachigkeit kann aber durchaus auch innerhalb einer Zielgruppe von „monoglots" existieren, wenn die Personen aus der Gruppe zwar jeweils nur eine einzige Sprache sprechen, diese jedoch von Person zu Person verschieden ist.

Eine intensive Beschäftigung mit den verschiedenen Facetten der Mehrsprachigkeit im IR findet bei den jährlich abgehaltenen *Conference and Labs of the Evaluation Forum* (CLEF) statt, die aus einer spezifisch cross-lingualen Aufgabenstellung der in den 1990ern gegründeten Initiative der *Text Retrieval Conference* (TREC) hervorgegangen sind, die ihrerseits auf den für die systemorientierte IR-Forschung wichtigen Cranfield Experimenten der 1960er Jahren aufbaut.[49] Dem von Stiller/Gäde/Petras 2013 formulierten Desiderat, die Anwendung von CLIR insbesondere in Gedächtnisinstitutionen zu untersuchen, widmet sich seit 2011 das im Rahmen von CLEF veranstaltete *ChiC Lab* (Cultural Heritgage in CLEF).[50]

Eine Grundfrage für das Retrieval in einem multilingualen Kontext ist, wie Nutzer_innen Zugang zu Dokumenten bekommen können, deren Sprache(n) sie nicht verstehen, und wie sie die Relevanz solcher Dokumente dennoch einordnen können. Eine Möglichkeit

[47] Siehe hierzu die präzise und bündige Definition von der Webseite des Cross Language Evaluation Forum (CLEF): „ If the user is utilising multiple languages to ask questions and accessing relevant information in multiple languages, they are utilising multiple language information retrieval (MLIR). In contrast, those who are asking questions in one language and accessing information in multiple languages are using the so-called cross-language information retrieval (CLIR). To simplify, MLIR is multilingual when it comes to both the query and information retrieval, while CLIR is multilingual only when it comes to facilitating access to relevant information." (http://www.clef-campaign.org/2006.htm, zuletzt geprüft am 11.03.2020).

[48] Vgl. Oard 2009, o.S.

[49] Zu TREC und CLEF siehe Petras 2013, S. 384. Zur Entwicklung der systemorientierten IR-Forschung seit den 1960ern siehe Ingwersen/Järvelin 2005, S. 1, 111. Seit 2001 gibt es zudem den interaktiven iCLEF track, der sich gezielt mit der Frage beschäftigt, ob Suchende relevante Dokumente innerhalb mehrsprachiger Ergebnislisten mit Hilfe von Übersetzungen gut identifizieren können; vgl. Oard 2009, o.S. Weitere relevante Fragen sind z.B., ob Suchende lernen können, effektive Anfragen zu stellen, und ob die Übersetzungen der Dokumente eine hinreichende Qualität haben, damit die Nutzer_innen ihre Inhalte verstehen; vgl. ebd.

[50] Vgl. Stiller/Gäde/Petras 2013, S. 93. Zur Auswertung der multilingualen digitalen Bibliothek Europeana in ChiC siehe außerdem Petras et al. 2012. Ein weiteres wichtiges Projekt auf europäischer Ebene zum Thema Multilingualität ist CACAO (Cross-Language Access to Catalogues and On-line Libraries), das zum Ziel hat, die cross-linguale Recherche innerhalb von OPACs und digitalen Bibliotheken in der EU zu ermöglichen; siehe hierzu u.a. Bosca/Dini 2009.

Sprachbarrieren zu überwinden bilden Übersetzungen der Dokumente selbst – entweder von Ausschnitten (Abstracts, ToC o.Ä.) oder der Volltexte. Gerade letzterer Fall ist jedoch sehr zeitaufwändig und kostspielig. Neben der Übersetzung der Dokumente selbst können aber auch die gestellten Suchanfragen oder die indexierten Terme übersetzt werden, sodass die Übersetzungsleistung bereits ins Systemdesign integriert wird.[51]

Was und wie viel zu welchem Zeitpunkt übersetzt wird, zieht jedoch viele Fragen nach sich, die nicht nur die Ebene der technischen Umsetzung betreffen, sondern auch Zeit- und Kostenfaktoren sowie grundsätzlichere Fragestellungen aus dem Bereich der Semiotik. So können Sprachen nie eins zu eins ineinander übertragen werden, fallen einige Begriffe der Ausgangssprache doch immer weiter oder enger aus als in der Zielsprache oder sind von Polysemie betroffen. Hinzu kommt die kontextuelle und pragmatische Dimension des Sprachgebrauchs, der auf rein semantischer Ebene nicht erfasst werden kann. Sprachen wie das Spanische, Portugiesische oder Englische, die weltweit in verschiedenen Ländern auf verschiedenen Kontinenten gesprochen werden, bilden schon in sich wenig homogene Entitäten und enthalten zahlreiche Bedeutungsvarianten je nach Herkunftsort der Sprecher_innen.[52] In der Folge erweist sich eine Eins-zu-Eins-Übersetzung zwischen den Sprachen als überaus schwierig. Dieses Äquivalenzproblem betrifft sowohl semantische als auch kontextuelle und strukturelle Aspekte, wie in den *Guidelines for Multilingual Thesauri* der IFLA ausgeführt wird.[53] So stellt neben der intra- und intersprachlichen Homonymie auch die kulturelle und strukturelle (Nicht-)Äquivalenz zwischen zwei oder mehreren Sprachen ein Problem dar; d.h. zum einen sind die Gebrauchskontexte betroffen und zum anderen die hierarchischen und assoziativen Beziehungen, in die ein Begriff in der jeweiligen Sprache eingebunden ist und die zwischen verschiedenen Sprachen nicht äquivalent sein müssen.

Auf der Indexierungsebene bieten hier nicht-symmetrische Thesauri sowie das Linking (oder Mapping) verschiedensprachiger Thesauri Möglichkeiten, mehrsprachige Zugangspunkte zu schaffen.[54] Neben solchen Crosskonkordanzen zwischen bestehenden Vokabularen kann auch die mulitlinguale Indexerweiterung durch die (dynamische) Übersetzung der Indexterme

---

[51] Vgl. hierzu Oard 2009, o.S. Der Autor geht auf drei verschiedene Ansätze bei der Übersetzung ein: 1) Übersetzung jedes einzelnen Terms in seinem Ursprungskontext, 2) Ermittlung der Häufigkeit des Vorkommens eines Terms und anschließende Übersetzung ohne die einzelnen Kontexte zu berücksichtigen, in denen er auftritt, 3) Ermittlung des Termgewichts jedes Terms und anschließende Übersetzung.

[52] Um nur ein kurzes Beispiel zu nennen: Suchen Nutzer_innen aus Argentinien nach dem Begriff „plata“, so meinen sie damit anders als Nutzer_innen aus Spanien nicht zwangsläufig Silber, sondern unter Umständen Geld. In dem von der Real Academia Española herausgegebenen Wörterbuch wird dieser abweichende Sprachgebrauch deshalb durch den Zusatz „Am.“ (Abkürzung für Lateinamerika) markiert; siehe: http://dle.rae.es/?id=TM1dBxX (zuletzt geprüft am 11.03.2020)

[53] Vgl. IFLA 2009, S. 3 sowie ausführlicher S. 11-18.

[54] Vgl. ebd., S. 2, 4 f., 16. Zum Linking/Mapping vgl. auch beispielsweise Kempf/Zapilko 2013 und Mayr/Petras 2008b. Ein zentrales Projekt in dieser Richtung ist das vom BMBF von 2004 bis 2007 geförderte und am GESIS angesiedelte Kompetenzzentrum Modellbildung und Heterogenitätsbehandlung (KoMoHe); siehe dazu Mayr/Petras 2008a.

mittels Transferwörterbüchern ein Weg sein, um den Bedürfnissen verschiedensprachiger Nutzer_innen gerecht zu werden. Andersherum können auch die Anfragen selbst übersetzt und so – im Falle eines einsprachigen oder einsprachig erschlossenen Bestands – auf das eingesetzte Indexierungsvokabular abgebildet werden.

Einige Projekte aus dem Bereich der automatisierten Inhaltserschließung berücksichtigen solche multilingualen Aspekte und streben eine Verbesserung des Retrievals etwa durch multilinguale Indexerweiterung an; die Einbindung multilingualer Thesauri wie z.B. EuroVoc bietet eine Möglichkeit hierzu.[55]

[55] Siehe: https://op.europa.eu/en/web/eu-vocabularies/th-dataset/-/resource/dataset/eurovoc (zuletzt geprüft am 11.03.2020).

Das Projekt AUTINDEX wurde am Institut für angewandte Informationsforschung der Universität des Saarlandes durchgeführt und bietet neben statistischen und computerlinguistischen Bearbeitungsschritten die Möglichkeit verschiedene, benutzerspezifische Thesauri und Klassifikationen einzubinden und automatisierte Indexierungsvorschläge zu generieren, die mittels Transferlexika in verschiedene Sprachen übersetzt werden können; vgl. hierzu Siegmüller 2007, S. 86-90 und Nübel/Schmidt 2003. Die von der Firma AGI entwickelte Anwendung IntelligentCapture, die die Software AUTINDEX einsetzt, die dazugehörige Recherchesoftware IntelligentSearch und die daraus hervorgegangene Plattform dandelon.com arbeiten im Bereich der (multilingualen) Katalogerweiterung und bieten eine übergreifende Suche in verschiedenen Bibliotheken sowie eine (auch mehrsprachige) Indexerweiterung, etwa durch die Einbindung (fremdsprachiger) Thesauri; vgl. hierzu verschiedene Publikationen von Manfred Hauer, u.a. Hauer 2004 und Hauer 2013 sowie Siegmüller 2007, S. 84-86. Zu der Anwendung von IntelligentCapture und dandelon.com in der Vorarlberger Landesbibliothek vgl. Rädler 2004. Maßgeblich auf Suchmaschinentechnologie und einer verteilten Suche basiert das in die 2004 initiierte Bielefeld Academic Search Engine (BASE) implementierte System FAST Data Search, das 2011 durch die Open-Source-Software Lucene/Solr abgelöst wurde. Neben der statistischen Gewichtung ermöglicht FAST die computerlinguistische Bearbeitung verschiedener Sprachen – laut Siegmüller konnten 2007 bereits 79 Sprachen erkannt werden; zu FAST vgl. Siegmüller 2007, S. 90-95. Ein Beispiel für den Einsatz computerlinguistischer Verfahren beim Retrieval in einer multilingualen Umgebung beschreibt Loth 2004, der die Verbunddatenbank des schweizerischen NEBIS-Verbunds untersucht, der mit dem im OSIRIS-Projekt entwickelten Recherche-System arbeitet.

2.3. Beschreibung der Ausgangslage: Sacherschließung am IAI[56]

Das Ibero-Amerikanische Institut Preußischer Kulturbesitz (IAI) beherbergt europaweit die größte Spezialbibliothek zu Lateinamerika, der Karibik und der iberischen Halbinsel. Zu Beginn der Durchführung dieser Studie 2017 umfasste der Bestand mehr als 1 Mio. gedruckter Monographien, über 23.000 E-Books, rund 3.800 laufende Abonnements von gedruckten Zeitschriften sowie ca. 6.000 E-Journals. Daneben halten die Sondersammlungen weitere Materialarten vor, darunter ca. 40.000 Tonträger und ca. 6.000 DVDs und Videos.[57] Diese verschiedenen Medien haben, dem Sammelgebiet des IAI entsprechend, einen Bezug zu Lateinamerika, der Karibik oder der iberischen Halbinsel, wobei der Sammelschwerpunkt auf den Geistes- und Sozialwissenschaften liegt. Sie sind zu einem großen Teil in den ibero-romanischen Sprachen (Spanisch und Portugiesisch) verfasst. Daneben finden sich in kleinerer Zahl Medien in Regionalsprachen (z.B. Katalanisch) oder indigenen Sprachen (z.B. Guaraní). Sowohl Übersetzungen von Primärwerken als auch die umfassende Sekundärliteratur zu den verschiedenen Aspekten von Sprache, Literatur, Kultur, Politik, Geschichte, Geographie usw. der entsprechenden Länder liegen auch auf Deutsch, Englisch, Französisch, Italienisch, Rumänisch und weiteren Sprachen vor. Die Nutzer_innen des IAI sind dementsprechend ebenfalls sehr multilingual aufgestellt. Genaue Zahlen zu dem Anteil nicht deutscher Nutzer_innen sowie den verschiedenen vertretenen Sprachgruppen lassen sich zwar nicht ermitteln, es kann jedoch vermutet werden, dass von den 2017 insgesamt 3.336 registrierten Nutzer_innen mindestens ein Drittel keine deutsche Staatsbürgerschaft besitzt.[58] Mehrsprachigkeit ist darüber hinaus aber ein deutlich umfassenderes Phänomen innerhalb der Nutzer_innenschaft, da auch deutsche Forschende in aller Regel mindestens eine der Sprachen des Sammelgebiets beherrschen.

Durch die Einbindung des Instituts in die Stiftung Preußischer Kulturbesitz (SPK) findet die Katalogisierung der Bibliothek kooperativ im Gemeinsamen Bibliotheksverbund (GBV) statt.

---

[56] Die Angaben zu Bestand, Sacherschließungspraxis und Aufbau des Online-Katalogs stammen aus internen sowie extern zugänglichen Quellen – wobei letztere der Webseite (https://www.iai.spk-berlin.de) entnommen werden können – oder wurden in persönlichen Gesprächen mit Kolleg_innen oder Vorgesetzten eruiert.

[57] Die genauen Angaben zu Bestand, Nutzung und weiteren Aspekten können auf der Webseite eingesehen werden, in einer Kurzfassung oder ausführlicher in dem auf der Seite hinterlegten PDF; siehe: https://www.iai.spk-berlin.de/das-iai/das-iai-in-zahlen-2017.html (zuletzt geprüft am 11.03.2020).

[58] Einer internen Statistik zu Folge lag 2017 allein der Anteil der registrierten Nutzer_innen mit einer anderen als der deutschen Staatsbürgerschaft bei 456 Personen. Allerdings befinden sich auch unter den 148 Nutzer_innen, die als Hochschulangehörige von Universitäten außerhalb Berlins geführt werden, nicht deutsche Personen. Gleiches gilt für die 896 Nutzer_innen, die nur einen Lesesaalausweis besitzen. Gerade bei dieser Gruppe kann der Anteil ausländischer Nutzer_innen als sehr hoch angenommen werden, da der Lesesaalausweis insbesondere von Personen genutzt wird, die per Fernzugriff aus allen Teilen der Welt auf die elektronischen Ressourcen des IAI zugreifen wollen, sowie von ausländischen Gastwissenschaftler_innen.

Bei der Katalogisierung in der Verbunddatenbank des GBV, die seit 2005 im PICA-Format geschieht, wurden im IAI für die Zugänge von 1994[59] bis 2016 flächendeckend Deskriptoren aus einem eigens von der Bibliothek des Instituts angelegten Thesaurus vergeben – im Weiteren als lokale Schlagworte (LSW) benannt. Unterschieden wird zwischen sieben Gruppen von Schlagworten: 1) Personenschlagworte, 2) Körperschaftsschlagworte, 3) Titelschlagworte, 4) Sachschlagworte, 5) Geographika,[60] 6) Zeitschlagworte, 7) Formalschlagworte.

Schlagworte von Zugängen vor 1994 wurden nicht aus den Zettelkatalogen in den OPAC übertragen und sind daher nicht online recherchierbar. Aufgrund des hohen zeitlichen Aufwands, den die intellektuelle Erschließung mit sich bringt, werden die seit 2016 erworbenen Monographien bis auf Einzelfälle nicht mehr verschlagwortet. Das Zugangsjahr ist an der Zugangsnummer ablesbar, die fast immer mit der Signatur nach Numerus Currens zusammenfällt.[61] Die Umstellung auf Numerus Currens erfolgte 1975 im Zusammenhang mit dem Wechsel von einer Freihandaufstellung zu einer Magazinbibliothek, beginnend mit der Signatur A 75/1. Der Altbestand, der im Außenmagazin in Friedrichshagen eingelagert ist, umfasst Zugänge, die vor diesem Datum erworben wurden und noch nach den alten Aufstellungssignaturen geordnet sind. Die Altsignaturen setzen sich aus einer übergeordneten Ordnungsgruppe für die geographischen Räume (Hauptsigel) und untergeordneten Sachsigeln zusammen.

Andere Medien werden weiterhin verschlagwortet, so etwa Zeitschriftentitel, DVDs und CDs sowie elektronische Ressourcen. Das Jahr 2016 bildet insofern keinen harten und absoluten Schnitt.

Der Thesaurus diente bereits als Inhaltserschließungsinstrument der Zettelkataloge. Mit der Umstellung auf eine elektronische Katalogisierungs- und Rechercheumgebung wurde dieser Ausgangsthesaurus bis 2016 sukzessive weiter angepasst. Veränderungen betreffen überwiegend die Schlagwortnormsätze von Personen oder Körperschaften. Die Sachschlagworte sind demgegenüber seit ihrer Erstellung seltener verändert oder angepasst worden und umfassen aktuell 4.676 Einträge (Stand 07.12.2017). Die Sprache des Thesaurus` ist überwiegend Deutsch; vereinzelt werden jedoch auch fremdsprachige Schlagworte vergeben oder in Form von Synonymen oder Eigenbezeichnungen integriert.

---

[59] 1994 wurde im IAI das Bibliothekssystem Urica eingeführt; 2005 erfolgte dann die Umstellung auf PICA; aktuell wird im CBS im PICA3-Format katalogisiert, im LBS werden die Katalogisate im PICA+-Format gespeichert. Als Software für die Katalogisierung im CBS wird die WinIBW eingesetzt.

[60] Hierbei wird weiter differenziert zwischen großen und kleinen Geographika, d.h. zwischen konkreten Orten wie z.B. Städten einerseits und geographischen Großräumen wie Ländern, Regionen, Kontinenten usw. andererseits.

[61] Unterschiede zwischen Signatur und Zugangsnummer treten v.a. bei mehrbändigen Werken auf, da die Signatur ausgehend von der Zugangsnummer des ersten erworbenen Bands gebildet wird und später erworbene Bände mit einer abweichenden Zugangsnummer dennoch die Grundsignatur des ersten erworbenen Bands übernehmen.

Von der Bibliothek des IAI wird keine systematische Klassifikation der Bestände vorgenommen. Die Schlagwortnormsätze sind jedoch größtenteils mit relationierten Notationen versehen, die den für die Altsignaturen verwendeten Sigeln entsprechen. Diese Notationen sind im Online-Katalog einsehbar, wenn das Schlagwort selbst angeklickt wird.[62]
Weitere im OPAC enthaltene Sacherschließungsinstrumente stammen aus der Übernahme von Fremddaten, die neben den Metadaten der Formalerschließung auch Schlagworte und Schlagwortketten aus der GND oder anderen Systemen umfassen sowie Notationen aus verschiedenen Klassifikationen, etwa der DDC oder der Basisklassifikation.[63] Seit der vollständigen Übernahme von PICA durch OCLC 2007[64] werden zudem Fremddaten aus dem WorldCat eingespielt. Da damit ein großer Teil der Fremddaten aus internationalen Nachweissystemen stammt, ist ein nicht unerheblicher Anteil der Erschließungselemente aus Fremddaten fremdsprachig, überwiegend auf Englisch.
Seit 2008 wird durch ein Online-Content-Verfahren (OLC) zudem auf Artikelebene formal erschlossen; dieses Verfahren wurde in den folgenden Jahren rückwirkend auf die Zeitschriftenzugänge bis ins Jahr 2000 zurück ausgeweitet. Bis 2000 wurden, nach Durchsicht durch die Fachreferent_innen, einzelne Artikel für die Katalogisierung ausgewählt und auch verschlagwortet. Sporadisch wurden allerdings auch noch zwischen 2000 und 2008 einzelne Artikel händisch ausgewählt, sodass sich auch in diesem Zeitraum vereinzelt verschlagwortete Beispiele finden können. Durch die flächendeckende maschinelle Erschließung auf Artikelebene durch das OLC-Verfahren ist eine Verschlagwortung der dadurch erzeugten Mengen an Datensätzen jedoch nicht mehr zu leisten.
Seit Anfang 2009 findet des Weiteren eine Kataloganreicherung statt, da Cover und ToC aller Bücher von der Bibliothek des IAI eingescannt und die ToC durch die Firma ImageWare indexiert und per OCR durchsuchbar gemacht werden. Diese Scanroutine wurde bereits 2000 in Form eines Current-Contents-Dienst[65] begonnen und als Anlass genommen, die Einzelauswahl von Artikeln für die Katalogisierung einzustellen; damals wurden jedoch nur

[62] Die Notationen sind jedoch nicht in der Suchmaske als Filterkriterium auswählbar, sodass anzunehmen ist, dass die Nutzer_innen keine Kenntnis von diesem Erschließungsinstrument nehmen. Zwar gibt es die Möglichkeit nach dem Filterkriterium „Lokale Systematik [LSY]“ zu recherchieren. Die Notationen werden hierdurch allerdings nicht aufgefunden, sondern stattdessen vereinzelt katalogisierte Sachgebiete. Das entsprechende Feld im PICA+-Format aus dem Schlagwortnormsatz im LBS wird zwar indexiert (Kategorie 045C), es wird jedoch keine Relation zu den Katalogisaten hergestellt. Da auf die Indexterme aus den Normsätzen der Schlagworte über eine indirekte Suche zugegriffen wird, ist diese Relation ausschlaggebend, um die Indexterme den Katalogisaten zuzuordnen. Zur Indexierung am IAI siehe genauer Kapitel 3.2.2.

[63] Seit 2003 ist das IAI als Teil der SPK Mitglied beim GBV. Allerdings begann die kooperative Katalogisierung und damit auch die systematische Fremddatenübernahme am IAI erst mit der Implementierung von LBS4 (PICA), die 2005 stattfand, und der Katalogisierung im CBS des Verbunds. Der GBV bezieht aktuell Fremddaten von der DNB, der GND, der ZDB, der LoC und Casalini; siehe: https://www.gbv.de/bibliotheken/verbundbibliotheken/02Verbund/02Verbundsystem/03Fremddaten/index (zuletzt geprüft am 11.03.2020).

[64] Siehe dazu: https://www.oclc.org/de/about/finance/mergers.html (zuletzt geprüft am 11.03.2020).

[65] Vgl. Castro Valle/Göbel/Lehmann 2005, S. 19.

die Images der Scans eingestellt. Mit der Umstellung 2009 wurden auch die seit 2000 angefertigten Images rückwirkend OCR-durchsuchbar gemacht.

Zusammenfassend heißt dies für die verschiedenen Bestandsgruppen:

- Die Erschließungsdaten für Medienzugänge vor 1994 sind nur über die Zettelkästen zugänglich.
- Medien, die zwischen 1994 und 2016 erworben wurden, sind mit Schlagworten im OPAC versehen, die online recherchiert werden können.
- Medien, die ab 2000 erworben wurden, sind systematisch auf Artikelebene nachgewiesen und über die aus der Kataloganreicherung extrahierten Indexterme (v.a. aus den eingescannten ToC) recherchierbar. Bei Auswahl nach Durchsicht, wie dies bis 2000 und teilweise noch bis 2008 praktiziert wurde, sind die Titelaufnahmen der Artikel vereinzelt zudem mit lokalen Schlagworten versehen.
- Medien, die ab ca. 2003 erworben wurden, sind durch die Verbundkatalogisierung mit Fremddaten angereichert.
- Monographien, die seit 2016 erworben werden, sind nur noch in Ausnahmefällen mit lokalen Schlagworten versehen; andere Medientypen wie Zeitschriftentitel, DVDs, CDs oder elektronische Ressourcen werden weiterhin verschlagwortet.

Besondere Faktoren im Hinblick auf die Verschlagwortung

Es gibt weitere Faktoren, die die oben dargelegten Zeiträume bezüglich der verschiedenen Erschließungsmethoden teilweise aufweichen und es erschweren, zeitlich eindeutig begrenzbare Parameter zu formulieren.
Zum einen betrifft dies solche Medien, die von der Retrokatalogisierung betroffen sind. Medien, die noch nicht retrokatalogisiert wurden, können im OPAC nicht aufgefunden und damit auch nicht in dieser Studie berücksichtigt werden (aktuell noch 13.290 Medien, Stand 23.03.18). Aber auch die Medien, die bereits retrokatalogisiert wurden, bilden einen Sonderfall, da sie in aller Regel lediglich mit dem für interne Auswertungszwecke angelegten Schlagwort „XXX (Import Retroprojekt)“ versehen werden und keine weitere inhaltliche Erschließung erfahren.

Der zweite Fall verzerrt die obige Darstellung insofern als eine Teilmenge der Zugänge von 2012 bis 2015 weder mit Schlagworten versehen noch über den Scan der ToC erschlossen ist. Dieser separat im Magazin aufgestellte Bestand ist formal bereits im OPAC nachgewiesen und bestellbar. Wird ein Medium aus diesem Bestand von den Nutzer_innen

bestellt, wird es aus dem Magazin ausgehoben und verschlagwortet; anschließend werden Cover und ToC eingescannt und ein Signaturschild angebracht. Es kann dann an die Nutzer_innen entliehen werden und wird nach der Rückgabe in den normalen Magazinbestand integriert. Die Zahl dieses bislang noch nicht vollständig eingearbeiteten Bestands liegt aktuell bei ca. 35.000 Monographien (Stand 28.03.2018).[66]

Merkmale des am IAI eingesetzten OPAC[67]

Am IAI wird ein OPAC eingesetzt, der inzwischen virtualisiert ist und von der VZG gehostet wird. Bei der Darstellung der Treffer kann zwischen einer chronologischen Sortierung (geordnet nach Erscheinungsjahr) und einer Sortierung nach Relevanz gewählt werden, wobei die chronologische Ordnung voreingestellt ist. Ebenfalls voreingestellt ist die einfache Freitextsuche („suchen [und]“) im Feld „[ALL] Alle Wörter“. Alternativ kann zu den Optionen „suchen [oder]“ oder „Index blättern“ gewechselt werden, und es können andere Suchfelder eingestellt werden (beispielweise „Alle Themen“, „Person, Autor“, „Titelanfang“ usw.). In der erweiterten Suche sind neben der Suche in mehreren Feldern folgende Filtereinstellungen möglich:

- Erscheinungsjahr
- Sprache
- Land
- Materialart

Die Treffer werden voreingestellt in einer Kurzliste von je 10 Treffern pro Seite angezeigt. Bei der Präsentation der Ergebnisse kann zu der Ansicht „Titeldaten“ gewechselt werden, in der nur jeweils ein Titeldatensatz angezeigt wird, oder zur „Suchgeschichte“. Das Retrieval im Online-Katalog des IAI erfolgt nach dem Prinzip des Exact Match und kann mittels Boole´scher Operatoren verfeinert werden.

[66] Für die noch nicht verschlagworteten Zeitschriften ist die Zahl schwerer zu ermitteln; sie liegt nach Schätzungen bei ca. 13.000 Titeln bzw. Einzelheften mit Stücktitel (Stand 28.03.2018). Das Problem bei der Ermittlung einer konkreten Zahl liegt darin begründet, dass nur die Gesamtaufnahmen der Zeitschriftentitel sowie vereinzelte Stücktitel verschlagwortet werden, die einzelnen Hefte hingegen nicht. Diese Fälle einzeln herauszufiltern wäre sehr komplex und zeitaufwändig gewesen, weshalb an dieser Stelle eine Grobschätzung nach Regalmetern vorgenommen wurde.

[67] Die Webadresse des OPACS lautet: https://www.iaicat.de/.

Abbildung 1: Suchmaske des OPAC des IAI

2.4. Forschungsfrage

Die vorliegende Arbeit hat zum Ziel, einen Beitrag zu der aktuell im Fokus der Öffentlichkeit stehenden Frage nach der Bedeutung kontrollierter Vokabulare zu stellen, in einer Zeit, in der vermehrt automatische Verfahren der Inhaltserschließung eingesetzt und (weiter)entwickelt werden.
In einem Retrievaltest sollen die Auswirkungen der Sacherschließung mittels eines lokalen, vorwiegend einsprachigen Thesaurus auf die Retrievalergebnisse innerhalb eines mehrsprachigen Bestands gemessen werden. Die zu Grunde gelegte Forschungsfrage lässt sich dabei wie folgt formulieren:

> *Welche Rolle spielen intellektuell vergebene, mehrheitlich deutschsprachige Schlagworte aus einem kontrollierten Vokabular bei der Suche in einem multilingualen Kontext?*

Primäres Ziel ist es festzustellen, ob und in welchem Maße die lokal vergebenen Schlagworte Anteil an der Generierung der Retrievalergebnisse zu ausgewählten Suchanfragen nehmen. Daraus lassen sich ganz allgemein folgende Bewertungskriterien ableiten: Bewertet wird, ob ein Trefferergebnis

- ausschließlich durch die lokalen Schlagworte erzeugt wurde,
- unter anderem durch lokale Schlagworte erzeugt wurde,
- gar nicht durch lokale Schlagworte erzeugt wurde.

Geprüft wird also, ob der getestete Suchterm in den lokalen Schlagworten indexiert wurde, und wenn ja, ob er *ausschließlich* dort indexiert wurde.
Um die Bedeutung der lokal vergebenen Schlagworte differenzierter beurteilen zu können, werden zwei gleich große Samples an Suchanfragen als Untersuchungsgrundlage dienen: ein deutschsprachiges und ein fremdsprachiges. Dadurch sollen Wert und Nutzen des deutschsprachigen kontrollierten Vokabulars für verschiedene Nutzer_innengruppen einander gegenübergestellt werden, womit auch die Forschungsfrage noch einmal differenzierter gestellt werden kann:

> *Welche Rolle spielen intellektuell vergebene, mehrheitlich deutschsprachige Schlagworte aus einem kontrollierten Vokabular bei deutschsprachigen gegenüber fremdsprachigen Suchen?*

Ausgehend von dieser Differenzierung der Forschungsfrage kann eine Vermutung bezüglich des Einflusses der Suchsprache auf den Anteil der lokalen Schlagworte formuliert werden:

> *Die Bedeutung der überwiegend deutschsprachigen lokalen Schlagworte variiert je nachdem, ob die Suchanfragen auf Deutsch oder in einer Fremdsprache formuliert werden.*

Den multilingualen Herausforderungen wird sich in dieser Arbeit von der Seite der Mehrsprachigkeit der Nutzer_innen genähert, die sich in der variierenden sprachlichen Verfasstheit der durch sie gestellten Suchanfragen äußert. Andere Aspekte der Multilingualität werden dagegen nicht berücksichtigt, beispielsweise sprachbezogene Voreinstellungen z.B. des OPAC-Interfaces oder eine Untersuchung der Interaktionen, die nach Auffinden der Dokumente vorgenommen werden, etwa nach welchen sprachlichen Kriterien Medien angesehen oder ausgeliehen werden. Ebenfalls nicht systematisch verfolgt werden kann an dieser Stelle die Multilingualität des Bibliotheksbestands und seiner Erschließung, die sich, wie in Kapitel 2.3. beschrieben, in viele weitere Facetten aufgliedern lässt. So können die Dokumente selbst in verschiedenen Sprachen verfasst sein – im IAI etwa stehen Quechua-Texte neben Übersetzungen von spanischer Lyrik ins Englische oder portugiesischsprachigen Filmen –, aber auch die beschreibenden Metadaten und die zugewiesenen Schlagworte können in verschiedenen Sprachen vorliegen – im IAI z.B. durch die Übernahme von LoC Subject Headings.

Um den Anteil der durch lokale Schlagworte gefundenen Dokumente zu ermitteln, werden 40 deutschsprachige und 40 fremdsprachige Suchanfragen an den Online-Katalog des IAI gestellt und die jeweils 10 ersten Trefferdokumente manuell daraufhin geprüft, in welchen Katalogfeldern der entsprechende Suchterm indexiert wurde. Differenziert wird dabei zwischen folgenden Variablen:

1) Titel
2) Weitere bibliographische Daten
3) Kataloganreicherung
4) Schlagworte (SW) aus Fremddaten
5) Lokale Schlagworte (LSW)

Diese fünf Variablen müssen für jedes gefundene Dokument überprüft werden, damit eindeutig festgestellt werden kann, ob die getesteten Suchterme noch in weiteren Katalogfeldern außer den lokalen Schlagworten indexiert wurden. Eine Fundstelle in den lokalen Schlagworten gibt noch keine Auskunft darüber, ob ein Dokument ausschließlich

durch lokale Schlagworte gefunden wurde, da noch weitere Variablen am Auffinden beteiligt sein können.

Die Relevanz der Trefferergebnisse wird in dieser Arbeit nicht untersucht; gerade bei den durch Schlagworte gefundenen Dokumenten kann sie jedoch als höher angenommen werden als bei solchen ohne Beteiligung kontrollierter Vokabulare.[68]

[68] Zu dieser Einschätzung vgl. auch Gross/Taylor/Joudrey 2015, S. 30. Die Autor_innen nehmen in ihrer Studie ebenfalls keine Relevanzbewertung vor, identifizieren eine solche aber als Desiderat; vgl. ebd., S. 30 f. Zum einen wäre es sehr heikel, ausgehend von einer Auswahl an Einwortanfragen aus einem anonymisierten Logfile realistische Aussagen über die Informationsbedürfnisse der Nutzer_innen zu treffen, die die Anfragen formuliert haben. Und zum anderen wäre eine Relevanzbewertung der Trefferergebnisse im Rahmen der Bearbeitungszeit der Masterarbeit, aus der diese Publikation entstanden ist, schlicht nicht zu leisten gewesen. Kelly hebt ganz grundsätzlich die Schwierigkeit hervor, geeignete Instrumente zur Messung der Informationsbedürfnisse von Nutzer_innen zu finden; vgl. Kelly 2009, S. 108 f. Auch Ingwersen/Järvelin unterstreichen die Ambiguität und Unvollständigkeit von Suchanfragen als Ausdruck von Informationsbedürfnissen der Nutzer_innen; siehe Ingwersen/Järvelin 2005, S. 114.

2.5. Gewählter Ansatz und Methode

In der vorliegenden Studie wurde ausgehend von echten Nutzer_innenanfragen an den OPAC des IAI ein Retrievaltest vorgenommen, der zum Ziel hat, den Anteil der lokal durch das IAI vergebenen Schlagworte am Auffinden der aufgefundenen Dokumente zu eruieren.
Es handelt sich dabei um eine deskriptive Evaluationsstudie,[69] in der die Häufigkeitsverteilung der Indexfelder, die für das Auffinden von Dokumenten zu insgesamt 80 verschiedensprachigen Suchanfragen verantwortlich sind, über fünf verschiedene Sucheinstiege ermittelt wird. Der Fokus liegt der Forschungsfrage folgend auf dem Anteil der durch lokale Schlagworte aufgefundenen Dokumente, wobei dem Aspekt der Multilingualität insofern Rechnung getragen wird als zwei verschiedensprachige Purpose Samples mit Suchanfragen der Nutzer_innen einander gegenübergestellt werden.[70]

Der Ansatz dieser Studie ist systemorientiert,[71] da der Fokus auf der Funktionsweise des OPACs liegt, konkreter auf dem Zustandekommen der Trefferergebnisse zu spezifischen Suchanfragen durch die Übereinstimmung von Dokument- und Anfragerepräsentation. Geschaut wird daher für jedes Trefferdokument, bei welcher der fünf Variablen der entsprechende Suchterm indexiert wurde. Dabei spielen insbesondere die von der VZG angewandten Indexierungsparameter eine entscheidende Rolle, die unter anderem die Zuordnung der Suchschlüssel zu den Indextermen regeln (siehe dazu Kapitel 3.2.2.).
Aus dem sehr komplexen Indexierungsverfahren und der Notwendigkeit einer manuellen Auswertung jedes einzelnen Trefferergebnisses folgt eine gewisse Gefahr im Hinblick auf die interne Validität des Verfahrens sowie die Reliabilität der Ergebnisse. So lässt sich nicht ausschließen, dass bei der Zuordnung der in den Indexaten aufgeführten Indexeinträge zu den Suchschlüsseln, die den Indexierungsparametern der VZG unterliegt, falsche Bewertungen getroffen wurden. Gleiches gilt für die Auswertung der erhobenen Daten, die in einer Excel-Tabelle erfasst und mit Hilfe von Filtern und Formeln für gezieltere Analysen extrahiert wurden. Hier wurde mit der größtmöglichen Sorgfalt vorgegangen und teilweise

[69] Zu den verschiedenen Arten von Studien, bei denen zwischen deskriptiven, explorativen und explanatorischen unterschieden wird, siehe Kelly 2009, S. 25 f. Ingwersen/Järvelin hingegen differenzieren deskriptive von komparativen oder explanatorischen Studien; vgl. Ingwersen/Järvelin 2005, S. 170.

[70] Durch die am Ende der Forschungsfrage formulierte Unterschiedshypothese bezüglich des Einflusses der gewählten Suchsprache könnte auch gezielt ein explanatorischer Ansatz verfolgt werden. Allerdings würde dies komplexere statistische Verfahren und Tests erforderlich machen, die hier nicht geleistet werden konnten und den Rahmen dieser Arbeit sprengen würden. Zum explanatorischen Ansatz vgl. Kelly 2009, S. 26.

[71] Zu den verschiedenen Ansätzen bei der Evaluation von Informationssystemen vgl. Petras 2013, im Hinblick auf systemorientierte Evaluationen insbesondere S. 371 f. Im Hinblick auf die historische Entwicklung des systemorientierten IR und seine Aufgabenfelder siehe Ingwersen/Järvelin 2005, S. 111-113. Für einen historischen Abriss über die Entwicklung der systemorientierten IR-Forschung siehe ebd., S. 120-122.

versucht auf unterschiedlichen Wegen die gewonnenen Daten zu überprüfen, um ihre Reliabilität zu gewährleisten.

Das Setting der Testläufe wurde durch verschiedene Kriterien kontrolliert, zum einen bei der Auswahl der Suchanfragen (siehe dazu Kapitel 3.1.2.) und zum anderen bei der Durchführung der Tests sowie der Auswertung der Trefferergebnisse (siehe Kapitel 3.2.1.). Der häufig geäußerte Einwand mangelnder Validität von Laborstudien aufgrund ihrer Abstraktion von den realen Suchbedingungen der Nutzer_innen[72] kann für die hier verfolgte Forschungsfrage allerdings insofern vernachlässigt werden als in dieser Studie das System selbst im Vordergrund steht: d.h. das von den Nutzer_innen genutzte Retrievalinstrument.
Ein Vorteil kontrollierter Testbedingungen liegt wiederum darin, gezielt die Wirkung einzelner Variablen messen zu können[73] – in diesem Fall den Anteil der durch die lokalen Schlagworte des IAI gefundenen Dokumente.
Inwieweit die Informationsbedürfnisse der Nutzer_innen durch die aufgefundenen Trefferdokumente erfüllt wurden, konnte dagegen nicht weiter untersucht werden; hierfür hätten z.B. Befragungen der Nutzer_innen zu ihrer Zufriedenheit oder eine Relevanzbewertung der Trefferergebnisse vorgenommen werden können.
Eine Annäherung an reale Suchbedingungen findet allerdings insofern statt als zum einen am laufenden System getestet wurde, d.h. in derselben Suchumgebung, die auch den Nutzer_innen zur Verfügung steht, und zum anderen, da die Testanfragen aus einem Transaction Logfile gewonnen wurden.[74] Sie drücken somit authentische Informationsbedürfnisse von Nutzer_innen mit unterschiedlichen Präferenzen bezüglich der gewählten Sprache ihrer Suchanfragen aus. Hier müssen jedoch klar die Grenzen von Logfiledaten als einer validen Basis für Aussagen über das Verhalten und die Motivationen von Nutzer_innen betont werden,[75] und allein eine eindeutige und verlässliche Zuordnung der in den URLs eines Logfiles gespeicherten Suchanfragen zu einer Sprache ist als problematisch anzusehen.[76]

---

[72] Zu dieser Kritik sowie weiteren Kritikpunkten an experimentellen Laborstudien in der IR-Forschung vgl. Ingwersen/Järvelin 2005, S. 4-9, 173 und 177; zu den Schwächen des dahinterstehenden systemorientierten IR-Ansatzes vgl. ebd., S. 186-189. Zu der Gegenüberstellung von Laborstudien gegenüber solchen in natürlichen Settings siehe Kelly 2009, S. 27 f. sowie insbesondere zum Aspekt der Validität S. 182. Kelly definiert Validität und Reliabilität folgendermaßen: „Validity is the extent to which methods and measures allow a researcher to get at the essence of whatever it is that is being studied, while reliability is the extent to which the method and measures yield consistent findings" (ebd., S. 181). Zu den Kriterien von Validität und Reliabilität siehe auch Greifeneder 2013, S. 261 und Fühles-Ubach/Umlauf 2013, S. 81.

[73] Vgl. hierzu Kelly 2009, S. 28 und Ingwersen/Järvelin 2005, S.173, 177.

[74] Zur Gewinnung und Aufbereitung der Daten aus dem Logfile siehe Kapitel 3.1.1.

[75] Zu den Grenzen und Begrenzungen von Logfiles und ihrer Analyse siehe das folgende Kapitel 2.6.

[76] Vgl. hierzu Gäde/Petras/Stiller 2010, S. 76. Dem Aspekt der Ambiguität bei der Erkennung der Query-Sprache widmen sich außerdem Stiller/Gäde/Petras 2010, o.S.

Die aus dem Logfile gewonnenen Anfragen dienten als Grundlage, um zwei gleich große Samples zu erstellen: auf der einen Seite ein deutschsprachiges und auf der anderen Seite ein fremdsprachiges. Hierbei wurden verschiedene Auswahlkriterien angewandt, sodass es sich bei den beiden Sets aus je 40 Suchanfragen um Purpose Samples handelt.[77]

Diese insgesamt 80 Suchanfragen wurden anschließend für einen Retrievaltest genutzt, bei dem der Anteil der durch den Sucheinstieg der lokal vergebenen Schlagworte aufgefundenen Dokumente ermittelt wurde. Als Cutoff-Wert wurden die ersten 10 Treffer festgelegt; dieser Wert bietet sich für den Test am OPAC des IAI insofern an als den Nutzer_innen voreingestellt die ersten 10 Treffer angezeigt werden.[78]

Für die statistische Auswertung der in den Testläufen gewonnenen Daten spielen die Häufigkeitsverteilungen von 80 getesteten Suchtermen in den Indexaten der zu ihnen aufgefundenen Dokumente eine Rolle. Im Fokus steht dabei das Vorkommen eines Suchterms in den indexierten Feldern der lokalen Schlagworte. Die Retrievalqualität sowie die Relevanz der gefundenen Treffer werden hingegen nicht systematisch bewertet, d.h. Kriterien wie Indexierungskonsistenz, Effektivitätskriterien wie Recall und Precision, Effizienz sowie die Zufriedenheit der Nutzer_innen finden in dieser Studie keine Berücksichtigung.[79]
In der anschließenden quantitativen Datenanalyse wurden ausschließlich diskrete Zahlen mit den Mitteln der deskriptiven Statistik ausgewertet.[80] Dabei wurden die Häufigkeitsverteilungen bezogen auf alle gefundenen *Dokumente* in Prozent normalisiert; außerdem wurde bezogen auf alle *Suchanfragen* die Abweichung vom Mittelwert der aufgefundenen Dokumente mit Hilfe der Standardabweichung errechnet.[81]

[77] Zum Vorgehen bei der Auswahl der Suchanfragen siehe Kapitel 3.1.2.

[78] Larson etwa gibt einen Durchschnittswert der von den Nutzer_innen pro Suche angesehenen Treffer von 9,1 an; vgl. Larson 1991, S. 209.

[79] Zu den möglichen Maßzahlen im traditionellen IR vgl. Siegmüller 2007, S. 22 f. und Ingwersen/Järvelin 2005, S. 173 f.; im Hinblick auf Maßzahlen im IIR siehe Kelly 2009, S. 103-105.

[80] Kelly definiert die Methode der deskriptiven Statistik folgendermaßen: „Descriptive statistics characterize variables; most notably these statistics describe central tendency and variation" (ebd., S. 135).

[81] Zu den statistischen Verfahren vgl. Bortz/Schuster 2010 sowie Schlittgen 1993, insbesondere Kapitel 7, S.136-153.

2.6. Literaturbericht

Für die hier behandelte Forschungsfrage wurde Literatur aus verschiedenen Bereichen des Information Retrieval (IR) konsultiert, wobei insbesondere Studien zur Bedeutung von kontrollierten Vokabularen für das Retrieval im Fokus stehen. Daneben spielt aber auch die Forschungsliteratur zur Bearbeitung und Analyse von Logfiles eine große Rolle, da auf diesem Wege die Suchanfragen der Nutzer_innen gewonnen wurden.

In der Forschung findet sich bereits seit den 1980ern eine Auseinandersetzung mit Transaction Logfiles oder anderen Formen von Logfiles.[82] Logfile-Analysen wurden und werden zu unterschiedlichen Zwecken eingesetzt, wobei in der Regel das Verhalten der Nutzer_innen von Informationseinrichtungen besser verstanden werden soll: ihre Vorgehensweisen und Schwierigkeiten bei der Recherche oder der Grad der Nutzung verschiedener Bibliotheksbestände – sei es aus einer intrinsischen Forschungsperspektive heraus oder stärker aus einer Managementperspektive, die daran interessiert ist, die Bibliotheksservices zu verbessern.[83]
Das in der Zeitschrift *Library Hi Tech* erschienene Themenheft zur Transaction Logfile Analysis (TLA) von 1993 fasst viele (noch immer) zentrale Aspekte dieser Methode zusammen und bietet einen guten Überblick sowohl über Chancen als auch Grenzen der TLA.[84] Schwierigkeiten im Zugang zu und bei der Lesbarkeit von Logfiles, wie sie etwa Flaherty konstatiert, sind auch der Autorin dieser Studie begegnet und betreffen v.a. die schiere Menge der in einem Logfile enthaltenen Daten, die damit verbundene mühsame Datenbereinigung sowie die für die Datenextraktion notwendigen Programmierkenntnisse.[85]

---

[82] Vgl. Fourie/Bothma 2007, S. 267. Kelly unterscheidet zwischen System Logging, Proxy Logging, Server Logging und Client Logging; vgl. Kelly 2009, S. 91-93.

[83] Vgl. Kaske 1993, S. 79, 81 und Blecic et al. 1998, S. 40. Ausführlicher zur Managementperspektive auf Logfile-Daten siehe Peters 1996. Sandore nennt als die möglichen Anwendungsfelder von TLA u.a. administrative Zwecke – etwa die statistische Auswertung im Hinblick auf Systemressourcen und Mitarbeitende –, technische Services, Management der Bibliotheksbestände, Systementwicklung und –verbesserung, Enduser-Feedback, IR und Nutzungsverhalten; vgl. Sandore 1993, S. 88-94. Insbesondere eine statistische Auswertung von personenbezogenen Daten z.B. der Beschäftigten ist aus datenschutzrechtlichen Gründen allerdings sehr heikel und stößt an andere Grenzen. Zu den wichtigsten Untersuchungsfeldern mit Blick auf das Suchverhalten der Nutzer_innen gehören laut Sandore problematische/erfolglose Suchen durch zu viele oder keine Treffer – insbesondere bei Schlagwortsuchen – sowie die Auswertung von Suchdauer und Menge der eingegebenen Suchterme; vgl. ebd., S. 89-91.

[84] Vgl. hierzu die Beiträge von Peters, Kurth, Kaske, Sandore und Flaherty von 1993. Einen Überblick über verschiedene Definitionen von TLA bieten Fourie/Bothma 2007, S. 266 f.

[85] Vgl. Flaherty 1993, S. 71, 77 sowie Kurth 1993, S. 99. Von den Autor_innen des Themenheftes wird deshalb eine Standardisierung der in Logs enthaltenen Daten gefordert; vgl. Sandore et al. 1993, S. 105 und Kurth 1993, S. 102 f. Priemer weist außerdem darauf hin, dass es auch bei den Programmen zur Analyse von Logdaten noch keine einheitlichen Standards gebe; vgl. Priemer 2004, S. 3. Kaske nennt als die vier zentralen Faktoren, die die Umsetzung einer TLA determinieren, die Verfügbarkeit der Daten, Mitarbeiter_innen mit entsprechenden Kenntnissen sowie Zeit und Geld; siehe Kaske 1993, S. 80, 83. Zu den Grenzen der TLA siehe insbesondere Kurth 1993 sowie Priemer 2004, S. 4 f.

Nach Kurth betreffen die Grenzen bzw. Begrenzungen von TLA v.a. vier Dimensionen: 1) die untersuchten Systeme, 2) die Nutzer_innen und ihre Suchen, 3) die Analyse der Logdaten, sowie 4) ethische und rechtliche Fragen.[86] Die einschränkenden Faktoren variieren je nachdem, um welche der Dimensionen es sich handelt, wobei der Autor auch betont, dass viele pragmatische Begrenzungen wie Zeit und Geld die echten, unumstößlichen Grenzen der TLA oft verschleierten.[87] Neben den bereits angesprochenen technischen Aspekten stehen v.a. die Nutzer_innen, ihr Suchverhalten und ihre Informationsbedürfnisse und damit kontextuelle und kognitive Grenzen im Vordergrund.[88] So können die in Logs gespeicherten Daten keine Auskunft darüber geben, wer mit welchem Informationsbedürfnis die Anfragen stellt und wie zufrieden die Person mit den Suchergebnissen ist, da die Logfiles nur die Struktur der Suchen abbilden können, nicht aber ihre Funktionen.[89]

Um diesen nutzerorientierten Aspekten stärker Rechnung zu tragen, war in den 1990er Jahren eine Tendenz zu beobachten, die darauf zielte, rein systembasierte und quantitative Methoden wie die TLA durch die Untersuchung kognitiver Aspekte mit Hilfe qualitativer Methoden zu ergänzen[90] – z.B. durch Fragebögen, Beobachtungen oder Selbstauskünfte von Nutzer_innen. Viele Autor_innen heben jedoch zugleich Wert und Nutzen der TLA als einer durch die quantitative Aufzeichnung sehr exakten und zugleich in der Analyse von Nutzer_innenverhalten nicht intrusiven Methode hervor – d.h. die Aufzeichnung bleibt für die Nutzer_innen unbemerkt und unterbricht ihren Rechercheprozess nicht – und appellieren darum häufig an eine Komplementarität der Methoden.[91]

Weiterentwicklungen der TLA, etwa zur Deep Log Analysis (DLA), beschäftigen die Forschung auch aktuell noch, und dies in steigendem Maße.[92] Der Aspekt der Multilingualität wurde laut Gäde/Petras/Stiller bislang in der TLA allerdings noch zu wenig berücksichtigt.[93]

---

[86] Vgl. Kurth 1993, S. 99 und ausführlicher S. 99-102.

[87] Vgl. ebd., 98 f.

[88] Vgl. ebd., S. 99 f.; Fourie/Bothma 2007, S. 273 f.; Priemer 2004, S. 5 sowie unter besonderer Berücksichtigung des Aspekts der Multilingualität Gäde/Petras/Stiller 2010, S. 77 f.

[89] Vgl. Kurth 1993, S. 98, 100 und Kelly 2009, S. 94.

[90] Vgl. hierzu z.B. Kurth 1993, S. 100, 102. Ingwersen/Järvelin beschreiben einen solchen "cognitive turn" im Hinblick auf das im IR eingesetzte systemorientierte Labormodell eingehender; siehe Ingwersen/Järvelin 2005, S. 3 f. und für eine ausführlichere Darstellung des kognitiven, nutzerorientierten Ansatzes die Kapitel 2 und 5.

[91] Vgl. Sandore 1993, S. 87, 89 f; Kurth 1993, S. 99 f.; Kaske 1993, S. 83; Fourie/Bothma 2007, S. 270 f., 279; Gäde/Petras/Stiller 2010, S. 70 sowie Griffiths/Hartley/Willson 2002 und Dawson/Williams/Gunter 2006. Priemer und Kelly heben außerdem positiv hervor, dass die TLA ohne zusätzlichen Zeitaufwand leicht und zuverlässig angewandt werden könne und es ermögliche, in authentischen Nutzungsumgebungen eine große Zahl von Nutzer_innen zu tracken; vgl. Priemer 2004, S. 4 und Kelly 2009, S. 94.

[92] Siehe hierzu etwa Fourie/Bothma 2007, S. 265. Fourie/Bothma sprechen in ihrem Beitrag nicht von TLA, sondern von Webtracking, da ihr Fokus auf dem Prozess des Sammelns und Speicherns der Nutzungsdaten liegt und nicht auf dem in Form von Transaction Logs fixierten Ergebnis; vgl. ebd. Die Autor_innen bieten zudem einen guten Überblick über die Forschungsliteratur bis 2006; vgl. ebd., S. 267-270. Priemer weist darauf hin, dass die Verfahren zur Speicherung von Logdaten und die Protokolle selbst oft synonym verwendet würden, unter den Begriffen Logfiles, Data-/Web-Mining oder User-Tracking; vgl. Priemer 2004, S. 1. Zur DLA vgl. u.a. Nicholas et al. 2006.

Eine zentrale Studie im Hinblick auf die Bedeutung von kontrollierten Vokabularen wurde von Gross/Taylor 2005 erstmalig durchgeführt und 2015 unter erweiterten Bedingungen von Gross/Taylor/Joudrey wiederholt.[94] In ihrer Studie von 2005 konnten die Autorinnen zeigen, dass kontrollierte Vokabulare für das Retrieval eine nicht unerhebliche Bedeutung haben und ca. ein Drittel der Retrievalergebnisse auf sie zurückgehen.[95] Für die Studie wurden Anfragen aus einem Transaction Log gewonnen; 186 der Anfragen wurden am PittCat, dem OPAC der Universität Pittsburgh, getestet, wobei nur englischsprachige Anfragen und Trefferergebnisse berücksichtigt sowie zu große Treffermengen und Nulltreffer ausgenommen wurden.[96] In einem zweiten Testlauf wurden nach der Einführung einer Kataloganreicherung durch ToC einige der bereits getesteten Suchanfragen erneut eingegeben. Auch hier zeigte sich, dass Schlagworte weiterhin eine große Rolle spielen; gleichzeitig wurde konstatiert, dass es zu einem Anstieg des Recall bei Einbußen an Precision durch irrelevante Treffer kam.[97]

In der jüngeren Studie von 2015 wurde die Testanordnung nach der Kataloganreicherung durch ToC und Abstracts wiederholt und um einen Testlauf in einer cross-lingualen Umgebung erweitert: Neben den Testläufen in englischsprachigen Beständen wurden in einem weiteren Testlauf nun auch fremdsprachige Dokumente berücksichtigt. Unter diesen veränderten Testbedingungen fiel der Anteil der Schlagworte an den aufgefundenen Dokumenten zwar geringer aus, es wäre aber in beiden Testläufen noch immer ungefähr ein Viertel der Dokumente ohne kontrollierte Vokabulare nicht gefunden worden.[98]

Als zentrale Vorzüge von kontrollierten Vokabularen benennen die Autor_innen u.a. die Verbesserung (Enhancement) der bibliographischen Datensätze, die Gruppierung von Synonymen und Sprachvarianten sowie semantische Disambiguierung, weiterführende Vorschläge durch Crossreferenzen, die Reduktion irrelevanter Treffer durch die gezielte Einschränkung und Spezifizierung der Suchen, die kontextuelle Einordnung der

---

[93] Vgl. Gäde/Petras/Stiller 2010, S. 72 f.

[94] Vgl. Gross/Taylor 2005 und Gross/Taylor/Joudrey 2015. Bereits 1995 befasste sich Taylor eingehender mit dem Thema der Subject Headings; siehe Taylor 1995. Für eine umfassende Darstellung der Forschungsliteratur zu diesem Thema siehe auch den ausführlichen Literaturbericht bei Gross/Taylor/Joudrey 2015, S. 3-24. Eine ältere Auseinandersetzung mit dem Thema findet sich z.B. bei Markey 1984.

[95] Vgl. Gross/Taylor 2005, S. 219. Die Autorinnen geben an, dass durchschnittlich 35,9% der durch die Anfragen gefundenen Dokumente ohne kontrollierte Vokabulare nicht gefunden worden wären; der Median liegt bei 30,2% und der absolute Anteil an allen gefundenen Dokumenten bei 35,4%.

[96] Vgl. ebd., S. 215 f. und 218 f.

[97] Vgl. ebd., S. 220.

[98] Vgl. Gross/Taylor/Joudrey 2015, S. 28-30. Für die englischsprachigen Dokumente lag der durchschnittliche Anteil der ausschließlich durch Schlagworte gefundenen Dokumente zu einer Suchanfrage bei 24,8% und der absolute Anteil an allen englischsprachigen Dokumenten bei 27,9%. Unter Einbezug auch fremdsprachiger Dokumente hingegen lag der durchschnittliche Anteil bei 27%, der absolute bei 27,7% und der Median bei 17,6%.

Trefferergebnisse und die thematische Erschließung auch von nicht textbasierten Medien.[99] Die Kosten, die durch die zeitaufwändige Erstellung und Pflege kontrollierter Vokabulare entstünden, zahlten sich dergestalt auf der Seite der Nutzer_innen aus, da sie dabei helfen könnten, deren Suchaufwand zu reduzieren.[100] Als mögliche Lösungsansätze, die in der Forschung gesehen werden, formulieren die Autor_innen die folgenden Optionen: 1) Komplementarität von kontrollierten Vokabularen und Keywords (Stichworte), etwa durch ein Nebeneinander von maschinell generierten Metadaten und menschlicher Kontrolle, 2) Tagging-Systeme, 3) Query Expansion durch die Anreicherung kontrollierter Vokabulare mit den von den Nutzer_innen eingegebenen Keywords, 3) Entwicklung von Tools, die die Nutzer_innen beim Umgang mit kontrollierten Vokabularen unterstützen, 4) Kataloganreicherung.[101]

Auf die Fragestellung von Gross/Taylor/Joudrey aufbauend widmet sich Garrett der Bedeutung von Subject Headings insbesondere in historischen Beständen in Volltextumgebungen.[102] Der Autor kommt zu dem Ergebnis, dass bei den von ihm untersuchten elektronisch verfügbaren historischen Textbeständen aus zwei verschiedenen Datenbanken (EEBO und ECCO) in 61,5% der Fälle die als Keyword eingegebene Phrase „east india company“ in den Subject Headings gefunden wurde; aus der darauf folgenden Auswertung von zwei kleineren Teilsamples von je 50 der zuvor gefundenen Dokumente ergibt sich des Weiteren, dass in 62% (EEBO) bzw. 60% (ECCO) der Fälle die Phrase *ausschließlich* in den Subject Headings gefunden wurde.[103] Der Autor kommt zu dem Schluss, dass die Volltextindexierung daher keinesfalls das Aus kontrollierter Vokabulare bedeute und diese gerade auch in Volltextumgebungen das Retrieval verbesserten, indem sie Bedeutungen innerhalb und über Sprachen hinweg disambiguierten und zusammenführten – eine Haltung, die Gross/Taylor/Joudrey zu Folge Konsens innerhalb der von ihnen untersuchten Forschungsliteratur seit 2004 sei.[104]

[99] Vgl. Gross/Taylor/Joudrey 2015, S. 31 und Gross/Taylor 2005, S. 223 f.

[100] Vgl. dazu den Literaturbericht bei Gross/Taylor/Joudrey 2015, S. 10 f.

[101] Vgl. ebd., S. 13 und ausführlicher Seite 13-19. Zu den unterschiedlichen Formen von Query Expansion bzw. Modification etwa durch Relevanz-Feedback oder kontrollierte Vokabulare siehe z.B. Ingwersen/Järvelin 2005, S. 140-150.

[102] Vgl. Garrett 2007. Zum Thema der Volltextsuche und ihrem Verhältnis zur menschlichen Kognition siehe außerdem den Beitrag von Garrett 2006. Weitere Untersuchungen zum Retrieval in Volltextumgebungen im Vergleich zum Auffinden von Dokumenten, die durch kontrollierte Vokabulare erschlossenen sind, finden sich beispielsweise bei Tenopir 1985 und McKinin et al. 1991.

[103] Vgl. Garrett 2007, S. 73 f.

[104] Vgl. ebd., S. 75 f. und Gross/Taylor/Joudrey 2015, S. 23.

Die Studie von Oberhauser/Labner zu einer OPAC-Erweiterung durch automatische Indexierung, die auf den MILOS-Projekten der 1990er Jahre aufbaut, kommt in verschiedenen Retrievaltests u.a. zu dem Ergebnis, dass Katalogisate, die mit Schlagworten versehen sind, innerhalb der aufgefundenen Dokumente sowohl vor als auch nach der automatischen Anreicherung des Indexes einen großen Anteil haben und sich beim automatischen Indexierungsverfahren als vorteilhaft erwiesen.[105] Ihr Resümee bezüglich der automatischen Indexierung fällt gleichsam positiv aus, da diese den Recall bei nur geringen Einbußen an Precision erhöht und den Anteil von Nulltreffern deutlich reduziert habe.[106]

Grundsätzlicher mit der Bedeutung verschiedener Sucheinstiege für das Retrieval befasst sich Wyly, wobei der Autor als Kriterium für das Retrieval als „erfolgreich" das Aufrufen von „circulation data" (d.h. Ausleih- oder Standortangaben) ansetzt und bemerkt, dass dieser Ansatz bislang kaum verfolgt worden sei.[107] Ausgehend von einer TLA kommt er zu dem Ergebnis, dass von allen aufgefundenen Titelsätzen, zu denen „circulation data" angefordert wurde, 29,9% durch Subject Headings gefunden wurden.[108]

Ein in der Forschung viel beachteter Aspekt ist die Gegenüberstellung von kontrollierten Vokabularen und (Titel-)Stichworten (Subject Headings vs. Keywords).[109] Durch die im Zusammenhang mit der Verbreitung verschiedener Best-Match-Modelle und das Relevance Ranking entstehenden, neuen Möglichkeiten für das IR in den 1990er Jahren[110] sowie die zunehmende Verbreitung von Suchmaschinentechnologie rückte die bei den Nutzer_innen beliebte Suche nach Keywords, die meist den Titeldaten entstammen, als eine vielversprechende Möglichkeit für das IR in den Blick.[111] Hintergrund ist die in verschiedenen Studien konstatierte Feststellung, dass die Nutzer_innen selten gezielt mit kontrollierten Vokabularen suchten und die dafür erforderlichen Suchstrategien in der Regel auch nicht ausreichend dominierten. Sehr häufig wird in diesem Zusammenhang die Studie von Larson von 1991 referiert, der v.a. gescheiterte Suchen (d.h. Nulltreffer) und Information Overload

[105] Vgl. Oberhauser/Labner 2003, S. 310-312.

[106] Vgl. ebd., S. 312.

[107] Vgl. Wyly 1996, S. 211 f., 233. Wyly knüpft nach eigener Aussage an kleinere Studien an, die von Gunnar Knutson 1986 und 1991 durchgeführt wurden, aufgrund ihres geringen Umfangs aber keine repräsentative Dimension hätten; vgl. ebd., S. 214 f.

[108] Vgl. ebd., S. 224, 233.

[109] Vgl. z.B. Rowley 1994; Tillotson 1995; Voorbij 1998 und Nowick/Mering 2003.

[110] Zur Entwicklung der verschiedenen Partial- bzw. Best-Match-Modelle siehe Ingwersen/Järvelin 2005, S. 117-119.

[111] Zu der jüngeren Forschungsliteratur zu diesem Themenkomplex siehe Gross/Taylor/Joudrey 2015, S. 4-7.

durch zu große Treffermengen für einen Rückgang in der Nutzung von Subject Headings verantwortlich macht.[112]

Furnas et al. sehen als Grund hierfür das „vocabulary problem“, mit dem auf die grundsätzliche Variabilität natürlichsprachiger Ausdrucksweisen abgehoben wird.[113] Damit benannt ist dasselbe grundsätzliche Verständigungs- und Kommunikationsproblem, das jedem Sprachgebrauch innewohnt: Nutzer_innen und Indexierer_innen müssen dieselben Begriffe für dieselben Sachverhalte wählen, damit eine Suche erfolgreich sein kann.[114] Insbesondere die „armchair“-Methode – d.h. die Festlegung kontrollierter Vokabulare durch eine_n einzige_n Indexierenden nach seiner/ihrer persönlichen Einschätzung, welcher Begriff einen Sachverhalt am exaktesten wiedergibt – ist den Autoren zufolge problematisch und wenig erfolgreich.[115] Die Lösung sehen Furnas et al. in der empirischen Rückkopplung der Terminologiewahl an den Sprachgebrauch der Nutzer_innen sowie die massive Ausweitung der Sucheinstiege, etwa durch die Extraktion von Termen aus Volltexten.[116] Dem aus Menge und Varianz der gewählten Begriffe resultierenden Problem der Ungenauigkeit müsse durch eine Disambiguierung mittels der Gewichtung der Begriffe nach Frequenz sowie interaktiven, lernenden Ansätzen begegnet werden.[117]

Garrett betont darüber hinaus das Problem des Sprachwandels, das bei der Suche in historischen Beständen besonders zum Tragen komme, und sieht deshalb trotz der Möglichkeit, in Volltexten nach Keywords zu suchen, kontrollierte Vokabulare als unverzichtbares Element, um orthographischen und semantischen Veränderungen innerhalb einer Sprache gerecht zu werden.[118] Als Forschungsdesiderat sieht der Autor insbesondere die Untersuchung abstrakter Konzepte, da diese anders als Eigennamen und Körperschaften in besonderem Maße von einem Bedeutungswandel über die Zeit hinweg betroffen seien.[119]

Auch Mann hebt die Bedeutung von kontrollierten Vokabularen gegenüber Keywords für die Forschung hervor, da sie einen systematischen Überblick böten, der gerade beim Umgang mit großen Sammlungen unverzichtbar sei, relevante von nicht relevanten Gebrauchskontexten segregierten und durch die Möglichkeiten des Browsing auch

[112] Vgl. Larson 1991, insbesondere S. 207-210 und 213 f. Wyly kommt zwar zu anderen Schlussfolgerungen als Larson, er konstatiert jedoch ebenfalls eine hohe Wahrscheinlichkeit, beim Subject Search mit kontrollierten Vokabularen Nulltreffer zu generieren; vgl. Wyly 1996, S. 221, 226-230.

[113] Vgl. Furnas et al. 1987, S. 964. Neben weiteren Autor_innen sprechen auch Ingwersen/Järvelin dieses Grundproblem an; vgl. Ingwersen/Järvelin 2005, S. 158 f.

[114] Vgl. Furnas et al. 1987, S. 964 f.

[115] Vgl. ebd., S. 965 f.

[116] Vgl. ebd., S. 968, 970.

[117] Vgl. ebd., S. 968-970. Auch Wyly plädiert für das Bereitstellen vielfältiger Sucheinstiege, insbesondere in Anbetracht steigender Mengen an Dokumenten; vgl. Wyly 1996, S. 233.

[118] Vgl. Garrett 2007, S.70, 75.

[119] Vgl. ebd., S. 75.

unerwartete Aspekte eines Themas ins Bewusstsein der Forschenden rücken könnten.[120] Er referiert darüber hinaus die vielfach konstatierte Diskrepanz zwischen den Suchstrategien der Forschenden gegenüber denen von Bibliothekar_innen oder anderen Informationsexpert_innen und appelliert für einen stärkeren Dialog zwischen beiden Welten.[121]

Zu anderen Schlussfolgerungen kommt Siegmüller – obgleich auch sie nicht für das Aufgeben intellektueller Erschließungsformen plädiert –, da ihrer Meinung nach Aufwand und Nutzen der Inhaltserschließung in keinem Verhältnis stünden und nicht alle Aspekte der Bestände erschlossen würden; sie appelliert darum an eine stärkere Integration der Sacherschließung bereits in die Suchanfragen, eine massive Erhöhung der Sucheinstiege, die Ablösung des Exact Match durch Ähnlichkeitssuchen und Relevance Ranking sowie hypertextbasierte Navigation, das Ausschöpfen aller medialen und technischen Möglichkeiten und eine stärkere Orientierung an den Nutzer_innen.[122]

Auch in den jüngeren Diskussionen um die inhaltliche Erschließung von Dokumenten durch Tags – also ein Bottom-up-Prinzip im Gegensatz zu dem Top-down-Prinzip kontrollierter Vokabulare – sind durchaus Tendenzen zu einer Kontrolle und Systematisierung der aus der Community eingebrachten Tags wahrzunehmen, im Sinne eines „tag gardening“ oder der „structured folksonomies“.[123]

Trotz des geäußerten Vorwurfs der Nutzer_innenunfreundlichkeit von Schlagworten galten thematische Suchen in kontrollierten Vokabularen (Subject Searches) in der Literatur bis zur Jahrtausendwende weiterhin als beliebtes, wenn nicht gar als *das* beliebteste Rechercheinstrument der Nutzer_innen bei der Suche nach Informationen.[124] Der von Larson 1991 konstatierte Rückgang von Suchen in kontrollierten Vokabularen hat folglich nicht zugleich einen Rückgang von thematischen Suchen zur Folge, sondern vielmehr deren Verlagerung in die Freitext- oder Titelsuche, wobei auch hier die Inhalte der kontrollierten Vokabulare mitdurchsucht werden.[125] In seiner vergleichenden Studie zur Recherche am Zettelkatalog und am OPAC einer indischen Spezialbibliothek kommt Sridhar zu dem Schluss, dass bei OPAC-Recherchen die am Zettelkatalog am häufigsten genutzten,

[120] Vgl. Mann 2005, S. 39 f.

[121] Vgl. ebd., S. 38, 40.

[122] Vgl. Siegmüller 2007, S. 51 f.

[123] Zu dieser Einschätzung vgl. Gross/Taylor/Joudrey 2015, S. 15.

[124] Vgl. Larson 1991, S. 197, 207; Sandore 1993, S. 90 und Sridhar 2004, S. 183. Sandore verweist insbesondere auf die Anfang der 1980er Jahre durchgeführten OPAC-Studien des CLR; vgl. Sandore 1993, S. 90. Larson zu Folge schwanken die Angaben zum Anteil der Schlagwortsuchen je nach Studie jedoch mitunter stark; vgl. Larson 1991, S. 197-199.

[125] Vgl. Gross/Taylor 2005, S. 213.

thematischen Suchen zu Gunsten von Suchen nach bereits Bekanntem – v.a. dem Titel der Dokumente – stark zurückgingen; der Autor führt dies auf eine mangelnde Indexierungsqualität sowie fehlende Informationskompetenz seitens der Suchenden zurück.[126] Besonders die in Zettelkatalogen sehr hilfreichen und reichlich vorhandenen Crossreferenzen gingen bei der Übertragung in Online-Kataloge häufig verloren, wenn sie nicht durch kontrollierte Vokabulare oder Online-Thesauri ersetzt würden.[127]
Ein Vorteil der Suche mit Keywords ist aus Nutzer_innensicht die Formulierung natürlichsprachiger Anfragen, ohne normierte Sucheinstiege kennen zu müssen. Auf Seiten der Informationssysteme bedeutet dieser Mangel an Normierung allerdings eine Herausforderung, da die zentralen Vorteile von kontrollierten Vokabularen damit ausbleiben (Zusammenführung von Synonymen, Herstellung hierarchischer Beziehungen zwischen Begriffen usw.) und aufgrund der sprachlichen Ambiguität der nicht normierten Stichworte viele irrelevante Treffer auftreten können. Die Forschung zur Verarbeitung natürlichsprachiger Anfragen im IR durch Natural Language Processing (NLP) hat sich seit den 1990er Jahren jedoch stetig weiterentwickelt und sich unter anderem intensiv mit den Möglichkeiten einer Verbesserung und Homogenisierung der Trefferergebnisse befasst.[128] Auch Entwicklungen aus der KI-Forschung prägen die Diskussionen im Bereich des IR, v.a. die Entwicklung wissensbasierter, lernfähiger Systeme, die durch Training neues Wissen ableiten können und damit von der Sprachoberfläche weiter auf die Ebene eines „Verstehens“ vordringen.[129]

Für die vorliegende Studie grundsätzlich relevant sind außerdem – meist aus der älteren Forschungsliteratur stammende – Untersuchungen zu den grundlegenden Schwierigkeiten von Nutzer_innen im Umgang mit OPACs,[130] da ein solcher auch am IAI eingesetzt wird. Zentrale Aspekte, die das Retrieval in einem OPAC beeinträchtigen, sind das Exact-Match-Prinzip und die Suche mit Boole´schen Operatoren, die den Nutzer_innen – so wie die Schlagwortsuche – häufig nicht bekannt oder wenig vertraut sind.

[126] Vgl. Sridhar 2004, S. 186-190.

[127] Vgl. ebd., S. 183.

[128] Zu den Möglichkeiten des NLP siehe Ingwersen/Järvelin 2005, S. 150-162. Viele dieser Aspekte wurden bereits in Kapitel 2.1. im Zusammenhang mit der Anwendung computerlinguistischer und begriffsorientierter Verfahren in der automatischen Inhaltserschließung angesprochen.

[129] Vgl. Siegmüller 2007, S. 26.

[130] Vgl. z.B. Scott/Trimble/Fallon 1995; Borgman 1986 und 1996 oder Schulz 1994. Auch Sridhar fasst die in der Forschung konstatierten gängigen Probleme von Nutzer_innen im Umgang mit OPACs zusammen; vgl. Sridhar 2004, S. 184 f. Speziell auf die Probleme beim Boole´schen Retrieval gehen Ingwersen/Järvelin ein; vgl. Ingwersen/Järvelin 2005, S. 119. Eine Übersicht über die Literatur zu den verschiedenen Suchstrategien der Nutzer_innen von OPACs findet sich bei Wyly 1996, S. 212-215.

Insbesondere der Aspekt „gescheiterter“ Suchen durch Nulltreffer oder zu große Treffermengen wurde in der älteren Forschungsliteratur viel beachtet und ausgehend von Logfile-Analysen untersucht.[131] Gründe für Nulltreffer sind u.a. Schreib- und Tippfehler, aber auch die Tatsache, dass das gesuchte Dokument im Bestand nicht vorhanden ist.[132]

Allerdings gibt es auch Stimmen, die grundsätzlich Kritik an der Definition einer Suche als „gescheitert“ üben. So kritisiert Wyly die von Larson gegebene Begründung für den Rückgang des Subject Search, nach der einerseits Nulltreffer und andererseits der Information Overload der Nutzer_innen durch eine zu große Treffermenge die Ursachen darstellten.[133] Wyly versteht den Suchvorgang der Nutzer_innen vielmehr als Kommunikationsprozess, der immer auch von „Scheitern“ geprägt sei; entscheidend sei jedoch, was aus diesem „Scheitern“ folge: So könnten Nulltreffer eine Abwandlung der Suche zur Folge haben, die einen Lernprozess beinhalte; und wenn bei großen Treffermengen gleich an erster Stelle das einschlägigste Referenzwerk zum gesuchten Thema auftauche, sei hier keinesfalls ein Information Overload anzunehmen; wichtig sei es darum für die Beurteilung des Erfolgs von Suchanfragen die Nutzer_innenperspektive stärker einzubeziehen.[134]

Siegmüller benennt als die Weiterentwicklungen im Bereich von OPAC-Systemen die feldübergreifende Suche, Tipps für weitergehende Recherchen bei Nulltreffern, kontextsensitive Hilfe, die Einbindung von Synonymen in den Schlagwortindex sowie die Umleitung von Anfragen zu einem passenden Schlagwort, das Dokumente zu dem Thema der Suche bündelt.[135]

---

[131] Vgl. hierzu Sandore 1993, S. 89 f.; Blecic et al. 1998, S. 48 und Flaherty 1993, S. 69. Ronthaler/Zillmann unterscheiden bei OPAC-Recherchen zwischen übergroßen Treffermengen, leeren Treffermengen und sehr kleinen Treffermengen; vgl. Ronthaler/Zillmann 1998, S. 1204. Untersuchungen zum Erfolg/Scheitern von Suchen bieten u.a. Peters 1989, Hunter 1991 und Zink 1991.

[132] Vgl. Sridhar 2004, S. 185; Blecic et al. 1998, S. 42 und Sandore 1993, S. 89 f.

[133] Siehe hierzu Fußnote 112.

[134] Vgl. Wyly 1996, S. 213 f.

[135] Vgl. Siegmüller 2007, S. 51.

3. Retrievaltest

3.1. Gewinnung und Aufbereitung der Daten

3.1.1. Aufbereitung des Logfiles

Für den in dieser Arbeit vorgenommenen Retrievaltest am OPAC des IAI wurden in einem ersten Schritt Purpose Samples aus authentischen Suchanfragen ausgehend von einem Transaction Logfile erstellt. Dieses Logfile stammt vom Apache Server, der vom GBV gehostet wird, und enthält die in den OPAC des IAI eingegebenen Suchanfragen. Das File umfasst die Einträge für den Zeitraum vom 28.02.2018 bis zum 21.03.2018. Ein größeres Logfile konnte aus technischen Gründen nicht zusammengestellt werden, da die VZG Files ab einer bestimmten Größe turnusmäßig löscht. Aus dieser Einschränkung ergaben sich verschiedene Schwierigkeiten im Hinblick auf die Datenlage, die in Kapitel 3.1.2. näher ausgeführt werden.

In dem Logfile mit 846.410 Einträgen sind enthalten:

- die IP-Adresse, von der aus die Anfrage gestellt wurde,
- Datum und Uhrzeit des Zugriffs,
- die gestellte Anfrage in URL-Form,
- der http-Status-Code,
- die Menge an übertragenen Bytes.[136]

Vor der Auswahl der Suchterme wurde das Logfile zunächst nach verschiedenen Kriterien bereinigt. So wurden nur die Hits beibehalten, die:

- die Datenbank 1 (DB1) durchsuchen: Dies ist die Datenbank des IAI; im Logfile enthalten waren zunächst auch DB2, DB3 und DB4, die zu anderen Einrichtungen der SPK gehören und auf dem Apache Server ebenfalls durch den GBV geloggt werden;[137]
- aus externen IP-Adressen stammen: Um dies zu gewährleisten, wurden die beiden vom IAI verwendeten IP-Adressen (Proxy und Server) aus dem File ausgenommen; diese beinhalten auch die im Lesesaal der Bibliothek aufgestellten PC-Terminals, die

[136] Nicht gespeichert werden die Antworten des Systems und die Zahl der gefundenen Treffer – beides Elemente, die die laut Flaherty standardmäßig in ein Logfile gehören oder gehören sollten; vgl. Flaherty 1993, S. 69, 74.

[137] DB2 = Staatliche Museen zu Berlin, DB3 = Staatliches Institut für Musikforschung, DB4 = Geheimes Staatsarchiv. Nach Ausschluss aller Datenbanken außer DB1 blieben noch 121.112 Einträge übrig.

von den Nutzer_innen benutzt werden können; die externen Anfragen stammen damit ausschließlich von den persönlichen Endgeräten der Nutzer_innen;[138]

- die Zeichenfolge „SRCHA&IKT=1016&SRT=YOP&TRM=“ enthalten: Damit wird sichergestellt, dass nur Suchanfragen im Logfile enthalten sind, die mit dem Suchschlüssel „Alle Wörter (ALL)“ und mit der voreingestellten Sortierung nach Erscheinungsjahr durchgeführt wurden;
- nach der Zeichenfolge „TRM=“ Buchstaben enthalten: Hierdurch werden Suchanfragen in Zahlenform ausgenommen;
- einen validen Statuscode 200 haben.[139]

Das derart bereinigte Logfile wurde anschließend in Excel importiert und in Spalten untergliedert. Es umfasst nach der Bereinigung noch 6.433 Hits (siehe Anhang 2). Die linke Spalte, die die externen IP-Adressen enthält, wurde zur Wahrung des Datenschutzes unmittelbar aus der Liste gelöscht, sodass die Anonymität der Anfragenden gewährleistet werden konnte.[140]

Durch diese Anonymisierung der Daten wurde folglich in keiner Form berücksichtigt, von welchen Personen die Suchterme eingegeben wurden. So ist es durchaus möglich, dass von den ausgewählten Suchtermen mehr als einer von derselben IP-Adresse aus eingegeben und damit (wahrscheinlich) von demselben/derselben Nutzer_in formuliert wurde, der/die unter Umständen einen einzigen thematischen Fokus bei der Suche verfolgt. Dies ist für die hier vorliegende Fragestellung jedoch nicht weiter relevant und wurde insofern ausgeglichen als versucht wurde, Begriffe aus unterschiedlichen Themenfeldern und Disziplinen auszuwählen – gleichwohl mit einem Fokus auf die in den Suchanfragen und im Bestand des IAI dominierenden Geistes- und Sozialwissenschaften. Für das deutschsprachige Sample konnte dies nur bedingt gewährleistet werden, da hier die geringe Datenmenge wenig Spielraum bei der Auswahl bot.

[138] Nach dem Ausschluss externer IP-Adressen blieben noch 101.371 Einträge übrig.

[139] Bei Beachtung der fünf hier genannten Kriterien, aber einer Umstellung der voreingestellten Sortierung der Ergebnisse von „Erscheinungsjahr“ auf „Relevanz“ (RLV) blieben nur 135 Anfragen übrig.

[140] Das Vorgehen dieser Studie wurde vor der Gewinnung der Logdaten mit der Datenschutzbeauftragten der SPK abgesprochen und ihr schriftlich geschildert. Außerdem wurde der örtliche Personalrat informiert, der dem Vorgehen ebenfalls zugestimmt hat. Zum Umgang mit Daten von Studienteilnehmer_innen vgl. z.B. Greifeneder 2013, S. 264-266.

3.1.2. Auswahl der Purpose Samples

Ausgehend von den Daten aus dem Logfile wurden zwei gleich große Samples gewonnen: einerseits ein deutschsprachiges und andererseits ein fremdsprachiges. Gerade bei den deutschsprachigen Suchanfragen an den OPAC war die Datenmenge allerdings eher gering. Ein Random Sample wäre aus diesem Grund nicht sehr sinnvoll und technisch schwer umsetzbar gewesen. Zudem erwiesen sich für die hier verfolgte Forschungsfrage verschiedene Kriterien bei der Auswahl der zu testenden Suchterme als sinnvoll, da diese von ihrer Form her so beschaffen sein sollten, dass sie zumindest potentiell als Schlagwort vorliegen könnten.

Die Auswahlkriterien für die Purpose Samples waren folgende:

- Die Suchterme wurden nicht in der englischsprachigen Suchmaske des OPAC eingegeben (erkennbar an der Syntax „LNG=EN“ in der URL der Hits aus dem Logfile), da es hier Abweichungen bei den bei der Indexierung eingesetzten Suchschlüsseln gibt; die englische Suchmaske wurde ohnehin lediglich bei 12 Anfragen benutzt, die für die Samples auch aus anderen Gründen nicht in Frage kamen.
- Die Hälfte, d.h. 40 Anfragen, sind in deutscher Sprache verfasst, die andere Hälfte, d.h. die verbleibenden 40 Anfragen, liegen in einer anderen Sprache als in Deutsch vor; die Fremdsprachen wurden auf Spanisch, Portugiesisch und Englisch beschränkt, da dies die am häufigsten verwendeten Suchsprachen am IAI sind.
- Der Suchterm ist nicht eindeutig als Personeneigenname oder Titel eines Dokuments identifizierbar.[141]
- Geographika wurden nur dann ausgewählt, wenn sich die fremdsprachige Bezeichnung von der deutschen unterscheidet.
- Der Suchterm ist wohlgeformt, d.h. orthographisch korrekt: Hierbei nicht berücksichtigt wurden allerdings die im Spanischen und Portugiesischen existierenden Diakritika und Akzente, da eine große Zahl der Nutzer_innen bei der Sucheingabe auf sie verzichtet. Außerdem ermöglichen Anfragen ohne Diakritka das Auffinden von Indextermen mit Diakritika, da diese mit Blick auf Groß- und Kleinschreibung sowie die in einem Dictionary-Verzeichnis hinterlegten

[141] Im Falle des für das deutschsprachige Sample ausgewählten Suchterms „flammenwerfer“ zeigte sich ausgehend von den aufgefundenen Dokumenten, dass hier vermutlich nach der deutschen Übersetzung des Romans *Los lanzallamas* von Roberto Arlt gesucht wurde. Diesen Zusammenhang habe ich aufgrund der übersetzten Version des Titels – zumal ohne Eingabe des Artikels – bei der Auswahl jedoch nicht hergestellt, weshalb ich auch diese Suchanfrage in die Ergebnisse miteinbeziehe.

Schreibvarianten behandelt werden (siehe dazu ausführlicher Kapitel 3.2.2.). Andersherum ist ein Matching zwischen Such- und Indexterm jedoch nicht möglich, d.h. beispielsweise Schreibweisen mit Akzent können nicht auf einen Indexterm ohne Akzent abgebildet werden. Für die Auswahl der Suchanfragen heißt das, dass auch Schreibweisen berücksichtigt wurden, bei denen Sonderzeichen fehlen, etwa beim Begriff „futbol“, der im Spanischen mit Akzent geschrieben wird („fútbol“); in diesen Fällen war es wichtig, dass dennoch eine Zuordnung zu einer der berücksichtigten Sprachen möglich war.

- Die Sprachzugehörigkeit entweder zum Deutschen oder zu einer der anderen drei vorrangig verwendeten Sprachen ist erkennbar:
  Dies war in vielen Fällen nicht eindeutig gegeben, insbesondere da die Groß- oder Kleinschreibung der Anfragen nicht als Distinktionsmerkmal gewertet werden kann; die Eingabe „transition“ etwa könnte genauso gut dem Deutschen wie dem Englischen entstammen, ebenso die Eingabe „war“. Solche Überschneidungen wurden nach Möglichkeit vermieden, in einigen Fällen ist aber durchaus Ambiguität in der Sprachzuordnung gegeben. Das Auswahlkriterium in diesen Fällen war, dass der Suchterm in jedem Fall auch zu der/den entsprechenden Sprache/n des Samples, dem er zugeordnet wurde, gehören könnte; d.h. für das deutsche Sample wurden im Einzelfall Begriffe verwendet, die im Deutschen definitiv vorkommen, gleichzeitig jedoch auch aus einer anderen Sprache stammen könnten. Anders verhält es sich mit eingedeutschten Eigennamen z.B. Sprachen, Musikstilen o.Ä. (etwa Guaraní oder Reggaetón), die aufgrund ihrer fremdsprachigen Herkunft für das fremdsprachige Sample benutzt wurden.
- Es wurden nur Einwortanfragen in Form von Substantiven ausgewählt: Einwortanfragen, um den Aufwand bei der manuellen Auswertung der Ergebnisse zu minimieren, und Substantive, da die Schlagworte im lokalen Thesaurus der Bibliothek des IAI in substantivierter Form vorliegen. Dabei wurden auch Suchanfragen berücksichtigt, bei denen nach dem Suchterm ein Leerzeichen eingeben wurde (in der URL erkennbar an dem auf den Begriff folgenden „+“-Zeichen), da in diesen Fällen unklar ist, ob der/die Nutzer_in tatsächlich vorhatte einen weiteren Suchterm einzugeben oder nur versehentlich ein Leerzeichen hinzugefügt hat.
- Es wurden keine sehr jungen Forschungsbegriffe berücksichtigt, da der Thesaurus des IAI seit 2016 nicht mehr systematisch gepflegt wird, insbesondere die Sachschlagworte nicht.

Die Auswahl eines deutschsprachigen Samples nach diesen Kriterien erwies sich als weitaus schwieriger als die der fremdsprachigen Suchterme, von denen sich deutlich mehr im zu Grunde gelegten Logfile finden ließen. Spanischsprachige Anfragen sind im Logfile besonders häufig vertreten, aber auch englischsprachige Suchterme liegen in großer Menge vor. Portugiesischsprachige Anfragen sind deutlich seltener zu finden; französisch-, italienisch- oder niederländischsprachige Eingaben konnten nur vereinzelt festgestellt werden.[142]

Ausgehend von einer Vorauswahl von rund 50 potentiell verwertbaren deutschsprachigen Einwortsuchen mussten im Verlauf der Tests noch einige Suchterme ausgeschlossen werden, da sie Nulltreffer generierten (siehe Dokumentation in Anhang 5) oder anderen der oben formulierten Kriterien nicht entsprachen. Aufgrund der schwierigen Datenlage musste dieses Sample von einer ursprünglich angestrebten Größe von 50 auf 40 Suchterme begrenzt werden. Das fremdsprachige Sample wurde in der Folge ebenfalls auf 40 Anfragen reduziert, um eine direkte Vergleichbarkeit der beiden Samples zu gewährleisten. Die ursprünglich geplante Zahl von 100 zu testenden Suchanfragen wurde damit auf 80 reduziert.

Da die Autorin selbst mit dem lokalen Thesaurus des IAI arbeitet und Medien verschlagwortet, ist ihr ein Teil der darin enthaltenen Schlagworte nicht unbekannt. Eine vollständige Objektivität bei der Auswahl der Suchterme war folglich nicht gegeben. Allerdings kann die Gefahr einer Verzerrung der Ergebnisse dennoch als gering angenommen werden. Zum einen, da insbesondere im Fall des deutschsprachigen Samples die geringe Datenlage sehr wenig Spielraum bei der Bevorzugung eines Suchterms gegenüber einem anderen ließ, und zum anderen, da die Deckungsgleichheit eines Suchterms mit dem kontrollierten Vokabular des IAI noch nicht automatisch bedeutet, dass das entsprechende Schlagwort auch bei den berücksichtigten Treffern vergeben wurde.[143]

[142] Eine Quantifizierung dieser Angaben, die durch eine manuelle Durchsicht des Logfiles gewonnen wurden, wäre nur mit einem erheblichen Aufwand möglich, da sich die Eingabesprache weder in der Excel-Tabelle gezielt filtern lässt noch, wie oben beschrieben, grundsätzlich ganz eindeutig bestimmen lässt. Bei den hier formulierten Einschätzungen handelt es sich folglich um nicht durch Zahlen belegbare Eindrücke.

[143] Das Problem der Indexierungskonsistenz wurde bereits in Kapitel 2.1. angesprochen.

3.2. Testaufbau und Durchführung

3.2.1. Kriterien der Testanordnung

Die in den Purpose Samples gesammelten Anfragen wurden erneut an den OPAC gestellt und die aufgefundenen Dokumente bis zu einem Cutoff-Wert von 10 manuell im Hinblick auf die Verteilung der entsprechenden Suchterme auf die Indexfelder geprüft. Ein solch detailliertes Verfahren war notwendig, da keine Möglichkeit zu einer automatisierten Auswertung der Indexfelder bestand und nur durch eine Prüfung des Indexes ermittelt werden konnte, in welchen Feldern außer den lokalen Schlagworten der jeweilige Suchterm indexiert wurde. Zwar können solche Dokumente, in denen der Suchterm (auch) in den lokalen Schlagworten gefunden wurde, durch den Suchschlüssel „and LSW + [Suchterm]“ ermittelt werden; allerdings bleibt damit die Frage ungeklärt, ob neben den Schlagworten noch weitere Indexfelder am Auffinden beteiligt waren. Aussagen darüber, wie viele Dokumente *ausschließlich* über die lokalen Schlagworte gefunden wurden, sind bei dieser Herangehensweise folglich nicht möglich, gerade dieser Aspekt ist für die hier verfolgte Forschungsfrage jedoch relevant.

Beim Retrievaltest wurden folgende Parameter für die Durchführung der Suche festgelegt:

- Die Suche erfolgt im Feld „Alle Wörter [ALL]“.
- Die voreingestellte Sortierung nach Erscheinungsjahr wird beibehalten; in der Folge kann es v.a. bei Suchanfragen, die große Treffermengen erzeugen, zu einer Überrepräsentation neuer Publikationen kommen.
- Die Eingabe der Suchterme wird durch den Suchschlüssel „LSW“ (Lokale Schlagworte) sowie eine Abfolge von Anfangszeichen ergänzt, die alle Buchstaben des Alphabets und alle Ziffern abdeckt:

lsw (a* | b* | c* | d* | e* | f* | g* | h* | i*| j* | k* | l* | m* | n* | o* | p* | q* | r* | s* | t* | u* | v* | w* | x* | y* | z* | 0* | 1* | 2* | 3* | 4* | 5* | 6* | 7* | 8*| 9*)

Dieser Suchschlüssel gewährleistet, dass nur Treffer, die mit lokalen Schlagworten versehen wurden, in der Ergebnisliste angezeigt werden.[144] Eine solche Einschränkung erscheint angesichts der hier verfolgten Forschungsfrage zielführend, da die Bedeutung der lokalen Schlagworte für das Retrieval von Dokumenten nur

[144] Schlagworte, die mit anderen Zeichen als in der oben angegebenen Klammer beginnen, sind im Thesaurus kaum enthalten (Ausnahmen können z.B. bei Titelschlagworten auftreten, die in <> eingefasst sind), und es kann davon ausgegangen werden, dass sie nicht alleine auftreten, d.h. ohne ein weiteres Schlagwort, das eines der Zeichen aus der Klammer enthält.

dann erhoben werden kann, wenn diese bei den aufgefundenen Dokumenten auch vorliegen. Andernfalls wäre eine verlässliche Angabe über die Zahl der nicht mit Schlagworten versehenen Dokumente am Gesamtbestand erforderlich gewesen und in die Ergebnisse einzubeziehen. Eine verlässliche Statistik liegt aufgrund der in Kapitel 2.3. beschriebenen Sonderfälle bei der Inhaltserschließung jedoch nicht vor.

- Die Eingabe erfolgt immer in Kleinschreibung, da bei der Verarbeitung der Suchanfragen durch das System die Unterscheidung von Groß- und Kleinschreibung keine Rolle spielt; ansonsten wird die Syntax der Originaleingabe aus dem Logfile befolgt, etwa im Hinblick auf Singular/Plural.
- Gesucht wird im laufenden System, d.h. es kann nicht ausgeschlossen werden, dass es je nach Suchzeitpunkt zu Abweichungen bei Menge und Anordnung der Trefferergebnisse kommt; zu jedem Testlauf wird deshalb das Datum der Durchführung notiert. Abweichungen innerhalb der geprüften Katalogisate sind dagegen verhältnismäßig unwahrscheinlich, da an diesen in der Regel keine Veränderungen mehr vorgenommen werden.

Bei der Auswertung der aufgefundenen Dokumente wurden folgende Einschränkungen vorgenommen:

- Anfragen, die Nulltreffer generieren, werden ausgenommen.
- Anfragen, die weniger als 10 Treffer generieren, werden berücksichtigt und es wird die entsprechende Zahl an gefundenen Dokumenten ausgewertet.
- Trefferergebnisse ohne lokale Schlagworte werden durch den bereits erwähnten „LSW“-Suchschlüssel ausgenommen. Eine Ausnahme bei der Auswertung der Ergebnisse bilden Medien, die nur mit einem einzigen Formalschlagwort erschlossen sind. Dies gilt in einigen Fällen für elektronische Ressourcen, die zwar prinzipiell thematisch erschlossen werden, in vielen Fällen aber zunächst nur ein Formalschlagwort („e-books“, „elektronische Zeitschriften“) pauschal zugeordnet bekommen. Auch bei CDs finden sich Fälle, in denen nur das Formalschlagwort „Tonträger“ vergeben wurde. Einen weiteren Sonderfall stellen ältere Titelsätze aus der Retrokatalogisierung dar, die in der Regel lediglich mit dem Schlagwort „XXX“ (Kommentar: „Import Retroprojekt“) versehen werden. Das Kriterium einer verbalen Inhaltserschließung ist in allen zuletzt genannten Fällen nicht gegeben. In Fällen, in denen sich ein solcher Treffer unter den ersten 10 Ergebnissen findet, wird dieser aus der Bewertung ausgeschlossen und dieser Ausschluss wird entsprechend in dem zu dem jeweiligen Suchterm gehörenden Dokumentationsdokument vermerkt.

- Die anderen 4 Variablen, die bei den gefundenen Dokumenten (bis zu einem Cutoff-Wert von 10) überprüft werden, müssen nicht zwangsläufig gegeben sein. D.h. Schlagworte aus Fremddaten und Elemente der Kataloganreicherung können vorliegen, müssen dies aber nicht, damit ein Treffer für die Auswertung berücksichtigt wird. Es wird jedoch für jedes Trefferdokument notiert, ob es über SW aus Fremddaten oder Kataloganreicherung verfügt, sodass nachträglich die Fälle, in denen alle Variablen vertreten sind, ermittelt und ausgewertet werden können.

3.2.2. Durchführung der Testläufe und Dokumentation

Einige Besonderheiten, die es bei der Durchführung und Dokumentation zu beachten gilt, ergeben sich aus den Indexierungsparametern, die von der VZG im Auftrag des IAI bei der Betreuung des OPAC angewandt werden, sowie den unterschiedlichen Systemen, die an der Indexierung beteiligt sind.[145] Bei den hier durchgeführten Tests mussten die Felder, die eine Fundstelle des jeweiligen Suchterms aufwiesen, in jedem Fall mit dem Suchschlüssel „ALL“ indexiert sein, um bei einer Suche im Feld „Alle Wörter“ auch tatsächlich aufgefunden zu werden.

Zunächst einmal muss festgehalten werden, dass zu den einzelnen Dokumenten kein Gesamtindexat vorliegt, das *alle* Indexterme enthält: d.h. die Indexterme aus den bibliographischen Daten, aus den Normsätzen sowie aus den verlinkten Elementen der Kataloganreicherung (z.B. gescannte ToC). In den direkt einsehbaren Indexaten zu den Katalogaufnahmen, die über eine versteckte Anzeige im OPAC hinterlegt sind,[146] sind lediglich die Katalogfelder zu finden, die aus dem CBS ins LBS übertragen wurden, mit ihren im LBS erstellten Indexeinträgen; des Weiteren eine Liste der Relationen zu den indexierten Normsätzen sowie Snippets[147] aus den Elementen der Kataloganreicherung, in denen der jeweils eingegebene Suchterm in seinem Kontext angezeigt wird. Die auf zweiter Ebene indexierten Terme, beispielsweise aus den Normsätzen der Schlagworte, sind in den Indexaten der Dokumente jedoch nicht zu sehen und müssen durch ein manuelles Öffnen des Normsatzes aufgerufen werden, da auf sie mit Hilfe der im Index definierten Relationen über eine indirekte Suche zugegriffen wird.[148] Für die Kataloganreicherung, das heißt v.a. die eingescannten ToC, liegt ein separater Volltextindex vor, der alle aus den Volltexten extrahierten Indexterme enthält. Diese Terme werden mit dem Suchschlüssel „TXT“ indexiert, der dem „Überindex ALL“ unterliegt. Sie sind aber, wie bereits erwähnt, nicht in der versteckten Anzeige aufgeführt.

Des Weiteren können Elemente aus dem CBS im LBS im PICA+-Format indexiert werden, ohne notwendigerweise in der Vollanzeige dargestellt zu werden. Dies gilt insbesondere für

[145] Die Auskünfte zu Aufbau und Funktionsweise des Indexes habe ich in Telefongesprächen und einem regen E-Mail-Verkehr mit Magdalena Roos von der VZG erhalten. Die Indexierungsparameter lassen sich einer Übersicht entnehmen, die ich ebenfalls von Frau Roos erhalten habe (siehe Anhang 3).

[146] Zu der versteckten Anzeige gelangt man, indem man unter das Icon klickt, das links von der Katalogaufnahme platziert ist und den Medientyp angibt. Die richtige Position erkennt man daran, dass sich der Cursor von einem Pfeil zu einer Hand verändert.

[147] Snippets sind kurze Textauszüge, in denen die Indexterme aus der Volltextindexierung im Kontext ihres Vorkommens angezeigt werden.

[148] Verknüpfungen zu Reihentiteln, zur Gesamtaufnahme bei einem mehrbändigen Werk, zum Zeitschriftentitel bei einem Zeitschriftenaufsatz sowie Umlenkungen zu dubletten Titelaufnahmen sind von dieser indirekten Suche dagegen nicht betroffen, da es sich hier um eine horizontale bzw. eine Family-Verknüpfung handelt. Die hiervon betroffenen Felder aus dem CBS sind folgende: 4160, 4180, 4241, 1698.

Normsätze aus dem Formalerschließungsteil, die bei der Darstellung in der OPAC-Anzeige für die Nutzer_innen nicht relevant erscheinen und deshalb dort nicht anklickbar sind. Diese Normsätze können damit Treffer generieren, die am OPAC nur durch eine Suche nach der PPN des entsprechenden Normsatzes geprüft werden können. Ein Beispiel hierfür findet sich beim Suchterm „kinderstadt“, der ausschließlich durch den Normsatz der Stadt Stuttgart im Formalerschließungsteil gefunden wurde, in dem dieser Term enthalten ist (PPN 106147811). Dieser Normsatz wurde über die indirekte Suche mitindexiert, in der OPAC-Anzeige jedoch nicht durchsuchbar dargestellt. Auch werden nicht zwangsläufig alle aus dem CBS in das LBS des IAI übernommenen Katalogfelder indexiert. Dies betrifft für die vorliegenden Daten nur sehr wenige Fälle, die den Indexierungsparametern der VZG entnommen werden können (siehe Anhang 3). Bei den hier durchgeführten Tests waren davon lediglich die PICA+-Felder 020F (in der OPAC-Anzeige als Kategorie „Inhaltsverzeichnis“ oder „Inhalt“ dargestellt), 220B, 046L und 046Q betroffen sowie verschiedene Unterfelder von Personennormsätzen oder Normsätze von Schlagworten aus dem lokalen Thesaurus oder aus Fremddaten. Bei der Kataloganreicherung gilt es ebenfalls die Indexierungsparameter zu beachten. So werden aktuell nur ToC indexiert, deren URL mit „https://www.gbv.de/dms“ beginnt.
Grundsätzlich gilt außerdem – dem Prinzip des Exact Match folgend – dass die Indexterme in ihrer vorliegenden Schreibweise indexiert werden; d.h. ein Suchterm kann ohne Trunkierung nicht gefunden werden, wenn er im Indexat in einer längeren Zeichenkette enthalten ist. Die Suchanfrage „mapuche“ im Singular beispielsweise wird in der indexierten Pluralform „mapuches“ nicht erkannt. Davon abzugrenzen sind Unterschiede bei der Darstellung im OPAC, etwa ob ein gefundener Term als Snippet angezeigt wird oder nicht. Wird ein Indexterm in einer abweichenden Schreibweise – etwa mit Akzent – gesucht, so wird er zwar aufgrund der Anwendung einer Konvertierungstabelle für Diakritika als deckungsgleich erkannt, es wird jedoch kein Snippet generiert. Snippets werden nur bei einer exakten Suche erzeugt. Auf die Erkennung abweichender Schreibweisen durch Diakritika (z.B. Umlaute oder Akzente) wird weiter unten noch einmal genauer eingegangen. Da der Prozess der Volltextindexierung unabhängig von der in einem zweiten Schritt durchgeführten OCR-Erzeugung in den hochgeladenen PDFs der ToC ist, kann es auch hier in Einzelfällen zu Unterschieden kommen, etwa wenn die OCR in dem den Nutzer_innen zugänglichen PDF nicht funktioniert.

Ausgehend von diesen Indexierungsparametern wurden die Testläufe mit den Suchtermen aus beiden Samples durchgeführt und dokumentiert. Die Arbeitsschritte der Durchführung waren folgende:

Im OPAC:

- Die im OPAC aufrufbare versteckte Anzeige des Indexes wurde geöffnet und nach dem jeweiligen Suchterm durchsucht.
- Alle hinterlegten Normsätze aus der Formal- und Sacherschließung wurden in der OPAC-Vollanzeige angeklickt und nach dem jeweiligen Suchterm durchsucht.
- Alle Links zu Elementen aus der Kataloganreicherung (Verlagsangaben, Rezensionen, gescannte ToC, Volltexte usw.) wurden geöffnet und nach dem jeweiligen Suchterm durchsucht.
- Bei all diesen Schritten wurde geprüft, ob die entsprechende Fundstelle mit „ALL“ indexiert wurde.

Im CBS:

- Waren im CBS Normsätze hinterlegt, die zwar ins LBS übertragen wurden, dort jedoch nicht in der Vollanzeige dargestellt werden, so wurden auch diese angeklickt und nach dem jeweiligen Suchterm durchsucht; fand sich der Suchterm darin, wurde die PPN des Normsatzes im OPAC aufgerufen und der Index des Normsatzes im PICA+-Format darauf geprüft, ob die entsprechende Fundstelle mit „ALL“ indexiert wurde.

Bei der Suche nach den Suchtermen in den Indexaten, in den Normsätzen sowie der Kataloganreicherung gilt es einige Besonderheiten im Hinblick auf die Verarbeitung durch das System zu beachten, etwa im Hinblick auf Diakritika wie Umlaute oder Akzente.
Bei der Verarbeitung der Suchanfragen an den OPAC des IAI sind Schreibvarianten zu den eingegebenen Termen mit Diakritika oder Sonderzeichen in der Regel auffindbar. Dies kann durch eine Verknüpfung zu einem Normsatz geschehen, in dem diese Schreibvarianten enthalten sind, oder durch eine Datei aus dem Dictionary-Verzeichnis, die dem Index hinterlegt ist (siehe Anhang 4). Während also das System beim Retrieval auch Schreibvarianten mit Diakritika auffindet und anzeigt, erlaubt die von mir nachträglich durchgeführte Suche nach dem jeweiligen Suchterm teilweise nur eine exakte Suche, etwa innerhalb des Dokumentationsdokuments in Word. Bei Wörtern, die in verschiedenen Schreibweisen existieren (innerhalb einer und zwischen verschiedenen Sprachen), habe ich deshalb nach dem Teil des Wortes gesucht, der unveränderlich bleibt. Außerdem wurde in den Fällen, in denen abweichende Schreibweisen im Index gefunden wurden, nachträglich am OPAC geprüft, ob das entsprechende Dokument mit der eingegeben Sprachvariante auch auffindbar ist. Hierzu wurde in das Suchfeld die PPN des Trefferdokuments mit dem Zusatz „and all [Schreibvariante]“ eingegeben. Nur in den Fällen, in denen sich die abweichende Schreibung in einem nicht indexierten Feld befand, konnte das Dokument bei

der Überprüfung nicht aufgefunden werden. Bei den im Deutschen üblichen Komposita war ein ähnliches Vorgehen notwendig, da diese vereinzelt in einer Schreibweise mit Bindestrich zu finden sind. Es wurde deshalb nach bloß einem der Bestandteile gesucht, um alle möglichen Schreibweisen auffinden zu können.

Die Dokumentation der Testläufe umfasste folgende Schritte:

- Von der ersten Seite der Trefferanzeige nach Eingabe mit und ohne lokale Schlagworte sowie von den ersten 10 Treffern (sofern es 10 oder mehr Treffer gab) wurden Screenshots erstellt und in einem Ordner unter der Bezeichnung des jeweiligen Suchterms abgelegt. Mussten Treffer ausgenommen werden, da nur ein Formalschlagwort vergeben wurde, wurde von ihnen ebenfalls ein Screenshot erstellt. In Fällen, in denen Treffer ab Position 11 aus der Trefferliste berücksichtigt werden mussten, wurde außerdem auch von der/den nächsten Seite/n der Trefferanzeige ein Screenshot abgelegt. Die Dokumentation der Suche ohne lokale Schlagworte dient dem statistischen Abgleich.
- Zu jeder Suchanfrage wurde ein Dokument angelegt, in dem das Datum der Testdurchführung notiert wurde und in das die LBS-Felder im PICA+-Format aus der versteckten Anzeige zu jedem berücksichtigten Treffer kopiert wurden. Die Fundstellen der Suchterme in den Indexaten wurden gelb hinterlegt, wobei an dieser Stelle noch nicht die Indexierungsparameter und das Exact-Match-Prinzip beachtet wurden. Welche Fundstellen tatsächlich mit „ALL“ indexiert wurden, kann Anhang 8 und 9 entnommen werden. Der Teil aus der versteckten Anzeige, in dem die den Indextermen zugewiesenen Suchschlüssel und Relationen aufgeführt werden, wurde nicht kopiert, da diese Zuordnungen dem Dokument mit den Indexierungsparametern der VZG entnommen werden kann (siehe Anhang 3). Die versteckte Anzeige kann darüber hinaus jederzeit am OPAC aufgerufen werden. Aus diesem Grund wurde auch darauf verzichtet, die Indexfelder der (lokalen sowie externen) Schlagworte zu kopieren, da auch diese in der versteckten Anzeige aufgerufen werden können und ihre Dokumentation sehr zeitaufwändig gewesen wäre.
- Außerdem wurde ein PDF mit den Titelaufnahmen der ersten 10 berücksichtigten Treffer aus dem CBS unter der Bezeichnung des jeweiligen Suchterms in dem dazugehörigen Ordner abgelegt (gekennzeichnet durch die Endung WinIBW); ausschließlich im CBS vorhandene Normsätze wurden ebenfalls als PDF im Ordner der zugehörigen Suchanfrage gespeichert.
- Für die Dokumentation der Verteilung der Suchterme auf die Indexfelder wurde eine Excel-Tabelle erstellt, die in der linken Spalte die eingegebenen Suchterme und die

Signaturen der maximal 10 berücksichtigten Treffer (bzw. die PPNs, sofern keine Signatur vorhanden ist, etwa bei elektronischen Medien) aufführt und in den anderen Spalten die Fundstellen der Suchterme bei den 5 für diese Studie festgelegten Variablen verzeichnet:

- **Titel:** Diese Variable umfasst sowohl Titel als auch Untertitel, Paralleltitel, Werktitel usw., ohne zwischen diesen genauer zu differenzieren.
- **Weitere bibliographische Daten:** Damit sind alle weiteren Felder aus dem Bereich der Formalerschließung gemeint, d.h. unter Umständen auch Normsätze von Personen, Körperschaften etc. sowie Inhaltsbeschreibungen oder –verzeichnisse, die in den bibliographischen Daten katalogisiert wurden.
- **Kataloganreicherung:** Damit gemeint sind Anreicherungen der Titelsätze durch Scans von ToC oder Links zu Rezensionen, Abstracts, Volltexten usw., d.h. jede Form der Weiterleitung zu externen Inhalten, die nicht im CBS selbst vorliegen; es wurde zudem notiert, wenn keinerlei Kataloganreicherung vorlag.
- **Schlagworte (SW) aus Fremddaten**: Diese Variable umfasst Schlagworte und Deskriptoren, die nicht vom IAI vergeben wurden – etwa solche aus der GND, die Subject Headings der LoC o.Ä.; Klassifikationen spielen aufgrund ihrer Zahlenförmigkeit in der Regel keine weitere Rolle, in Einzelfällen werden aber auch die Bezeichnungen der Haupt- und Unterklassen zu den Notationen angegeben und indexiert; es wurde zudem notiert, wenn keinerlei kontrolliertes Vokabular aus Fremddaten vorlag.
- **Lokale Schlagworte (LSW):** Hiermit sind die aus dem lokalen Thesaurus des IAI vergebenen Schlagworte gemeint; das entsprechende Schlagwort wurde in einem Bemerkungsfeld notiert.

Für die Titelvariable und die weiteren bibliographischen Daten wurde jeweils das Feld aus dem Index des LBS notiert. Das Kriterium der Zuordnung zu den hier untersuchten Variablen ist allerdings die Darstellung in der OPAC-Anzeige. Zwar ist die Zuordnung von Indexfeldern zu den Kategorien der OPAC-Anzeige in den allermeisten Fällen regelhaft, wenige Ausnahmen finden sich jedoch. So gibt es einige wenige Fälle, in denen das Anmerkungsfeld auch Informationen zum Titel enthält, die dann mit einem entsprechenden Suchschlüssel, etwa „TIT“ (Titel), „TAF“ (Titelanfang) o.Ä. indexiert wurden. Allerdings werden bei der Indexierung in der Regel gleich mehrere Suchschlüssel vergeben, sodass der Suchschlüssel oft nicht als eindeutiges Zuweisungskriterium fungieren kann. Deshalb wurde für die vorliegende Arbeit die Entscheidung getroffen, die Kategorien aus der OPAC-Anzeige als Kriterium für die Zuordnung zu der Variablen „Titel“ anzuwenden.

- Daneben wurden in einer separaten Tabelle die Treffermengen der jeweiligen Suchanfragen vermerkt – zum einen die Gesamttreffermenge und zum anderen die Menge an Treffern unter der Einschränkung, dass lokale Schlagworte vorliegen (siehe den Suchschlüssel aus Kapitel 3.2.1.).
- Außerdem wurden in einer weiteren Excel-Tabelle solche Fälle dokumentiert, in denen sich der eingegebene Suchterm mit Normsätzen aus dem kontrollierten Vokabular des IAI oder solchen aus Fremddaten deckt.

3.2.3. Auswertung und Ergebnisse der Testläufe

Zunächst ist festzuhalten, dass die Ergebnisse aus den Testläufen unter Vorbehalt und mit gewisser Vorsicht zu behandeln sind. Auch wenn die Tests nach bestem Wissen und Gewissen durchgeführt wurden, bergen die manuelle Durchsicht der Ergebnisse – zumal in unterschiedlichen Systemen – und die anschließende, ebenfalls manuelle Dokumentation eine hohe Fehleranfälligkeit. Hinzu kommen die Besonderheiten bei den Indexierungsparametern, wie in Kapitel 3.2.2. geschildert, die ebenfalls bei jedem Testlauf beachtet werden mussten.
Die beiden Samples (deutschsprachig und fremdsprachig) aus je 40 Suchanfragen wurden zum einen jeweils einzeln ausgewertet; zum anderen wurde der Anteil der durch LSW gefundenen Dokumente über die Gesamtmenge von 697 Dokumenten aus allen 80 Anfragen ermittelt.

Bei der Auswertung der gesammelten Daten wurden die für diese Studie relevanten Zahlen mit Hilfe von Filtern aus den verschiedenen Excel-Tabellen extrahiert. Dafür wurden die Spalten benutzt, in denen durch die Eintragung der Ziffer „1“ (grün hinterlegt) vermerkt wurde, ob der jeweilige Suchterm durch die entsprechende Variable gefunden wurde. Im Fall der Kataloganreicherung und der SW aus Fremddaten wurde – ebenfalls durch die Ziffer „1“ (in diesem Fall rot unterlegt) – in der entsprechenden Spalte markiert, wenn diese Variable im Katalogisat nicht vorhanden war.

Anzahl gefundener Dokumente pro Suchanfrage

Bezüglich der Anzahl der gefundenen Dokumente pro Suchanfrage bei einem Cutoff-Wert von 10 ist festzustellen, dass die Anfragen des deutschsprachigen Samples insgesamt weniger Ergebnisse erzielten und 35% weniger Treffer mit 10 oder mehr Ergebnissen hervorriefen als das fremdsprachige Sample (siehe Abb. 2). Die Ergebnismengen schwankten zwischen 0 und 10, wobei Nulltreffer für die Studie nicht berücksichtigt und aus dem Sample ausgenommen wurden (siehe dazu die Dokumentation in Anhang 5). Im fremdsprachigen Sample erzielte lediglich die Suchanfrage „quietismo“ weniger als 10 Ergebnisse. Der Mittelwert der berücksichtigten Dokumente pro Suchanfrage liegt im fremdsprachigen Sample bei 9,85 Dokumenten, während er beim deutschsprachigen bei nur 7,58 Dokumenten liegt (siehe Abb. 2). Für beide Samples zusammengenommen beträgt der Mittelwert der zu den Suchen aufgefundenen und berücksichtigten Dokumente 8,71.

Abbildung 2: Anzahl der Dokumente pro Suchterm im Vergleich beider Samples[149]

Anzahl Dokumente pro Suchterm	Suchterme mit entsprechender Dokumentzahl Sample deutschsprachig	%	Suchterme mit entsprechender Dokumentzahl Sample fremdsprachig	%	Differenz in %
10 oder mehr	25	62,50	39	97,50	-35,00
8	2	5,00		0,00	5,00
7	1	2,50		0,00	2,50
5	1	2,50		0,00	2,50
4	2	5,00	1	2,50	2,50
3	3	7,50		0,00	7,50
2	2	5,00		0,00	5,00
1	4	10,00		0,00	10,00
Gesamtsumme	303	100	394	100	
Mittelwert	7,58		9,85		
Nicht berücksichtigte Nulltreffer	5		0		

Diese Diskrepanz zwischen den beiden Samples bestätigt sich, wenn man von den realen Ergebnismengen pro Suchanfrage ausgeht, d.h. wenn Treffermengen oberhalb des Cutoff-Wertes von 10 berücksichtigt werden.

So liegt der Mittelwert des deutschsprachigen Samples für solche Treffer, die dem eingegebenen Suchschlüssel gemäß in jedem Fall mit LSW versehen sind, bei 2.246,05 Dokumenten (siehe Anhang 13) gegenüber einem Mittelwert von 3.433,63 (siehe Anhang 14) Dokumenten im Falle des fremdsprachigen Samples. Ohne den einschränkenden Suchschlüssel beträgt der Mittelwert für das deutschsprachige Sample 2.729,83 (siehe Anhang 13) und für das fremdsprachige 5.708,75 (siehe Anhang 14). Für beide Samples zusammengenommen liegt der Mittelwert aufgefundener Dokumente mit LSW bei 2.839,84 und für die Suche ohne diese Einschränkung bei 4.219,29 gefundenen Dokumenten (siehe Anhang 12).[150]

[149] Für eine schnellere Orientierung der Leser_innen werden die Angaben zu den Einzelsamples farblich differenziert. Angaben zum deutschsprachigen Sample werden orangefarben unterlegt, solche zum fremdsprachigen blau.

[150] Der Aspekt der „gescheiterten“ Suchen spielt für die Forschungsfrage dieser Arbeit keine weitere Rolle. Die Zahlen zu den durch die Anfragen erzeugten Treffermengen legen jedoch nahe, dass dies ein interessanter Aspekt für weitergehende Untersuchungen sein könnte. Sowohl sehr kleine Treffermengen (unter 10) oder Nulltreffer sowie sehr große Treffermengen können aus Sicht der Nutzer_innen ein Hindernis bei der Auswertung der Suchergebnisse darstellen. Die Mittelwerte zu den hier vorgenommenen 80 Anfragen zeigen, dass die Mehrheit der Anfragen sehr große Treffermengen erzielt.

Als Folge dieser Schwankungen bei der Menge an Trefferergebnissen wurde der Grad der Beteiligung der LSW am Auffinden der Dokumente unter zwei verschiedenen Perspektiven ausgewertet:

1) Normalisiert in % bezogen auf die Gesamtmenge der gefundenen Dokumente a) des deutschsprachigen Samples (303 Dokumente), b) des fremdsprachigen Samples (394 Dokumente), c) beider Samples zusammengenommen (697 Dokumente).
2) Studentisiert als Abweichung vom arithmetischen Mittel der durch LSW aufgefundenen Dokumente bezogen auf die Gesamtmenge der Suchanfragen a) des deutschsprachigen Samples (40 Anfragen), b) des fremdsprachigen Samples (40 Anfragen), c) beider Samples zusammengenommen (80 Anfragen).[151]

Betrachtung beider Samples zusammengenommen

Der Anteil der durch LSW gefundenen Dokumente pro Suchanfrage und der Grad der Beteiligung der LSW an ihrem Auffinden lässt sich detailliert Anhang 12 entnehmen. Die in Kapitel 2.4. formulierten Arten der Beteiligung der LSW am Auffinden lassen sich für beide Samples zusammengenommen folgendermaßen zusammenfassen:

Abbildung 3: Beteiligung LSW am Auffinden der Dokumente in beiden Samples zusammengenommen

Art der Beteiligung LSW	Anzahl gefundener Dokumente	%
Ausschließlich durch LSW	133	19,08
Durch LSW sowie mindestens eine weitere Variable	39	5,60
Nicht durch LSW	525	75,32
Gesamtsumme	697	100

[151] Die pro Anfrage gefundenen Dokumente (=n) wurden hierfür durch folgende Formel studentisiert:

$$z_i = \frac{x_i - \bar{x}}{\sqrt{\frac{1}{n}\sum_k (x_k - \bar{x})^2}}$$

z_i: Studentisierter Wert; x_i: Originalwert; $\bar{x}$: Arithmetischer Mittelwert; n: Anzahl; i, k: Laufvariablen.

Dieses Vorgehen wird auch als Z-Transformation bezeichnet; vgl. hierzu Bortz/Schuster 2010, u.a. S. 71, 586.

Werden, bezogen auf die Gesamtmenge aller Dokumente, die Teilmengen an Dokumenten betrachtet, die ausschließlich über eine einzige der 5 Variablen aufgefunden wurden, so ergibt sich folgende Rangfolge der verschiedenen Sucheinstiege (siehe Abb. 4):

Kataloganreicherung:	299 Treffer (42,9%)
LSW:	133 Treffer (19,08%)
SW aus Fremddaten:	73 Treffer (10,47%)
Bibliographische Daten:	51 Treffer (7,32%)
Titel:	46 Treffer (6,6%)

Abbildung 4: Ausschließlich durch eine einzige Variable gefundene Dokumente in beiden Samples zusammengenommen

Variable	Anzahl gefundener Dokumente	%
LSW	133	19,08
SW aus Fremddaten	73	10,47
Kataloganreicherung	299	42,90
Weitere bibliographische Daten	51	7,32
Titel	46	6,60
Summe[152]	602	86,37
Rest	95	13,63
Gesamtsumme	697	100

Vergleich der Samples

Zwischen den beiden getrennt ausgewerteten Samples sind deutliche Unterschiede in Bezug auf den Anteil der verschiedenen Sucheinstiege am Auffinden der Dokumente erkennbar. Eine detaillierte Übersicht über die Ergebnisse der einzelnen Anfragen kann für das deutschsprachige Sample Anhang 13 entnommen werden und für das fremdsprachige Anhang 14.

Der Grad der Beteiligung der LSW beim Auffinden der Dokumente lässt sich für jedes einzelne Sample folgendermaßen zusammenfassen (siehe Abb. 5 und 6):

[152] Die prozentualen Anteile der Zwischensummen wurden in allen Tabellen ausgehend von den diskreten Zahlen der gefundenen Dokumente berechnet und nicht durch Addition der prozentualen Anteile der Variablen, da es hier durch die Rundung der Werte auf die zweite Kommastelle zu Unschärfen hätte kommen können. Die Gesamtsumme hingegen wurde durch Addition aller Werte überprüft, auch bei den Prozentangaben. Kam es hierbei durch die Rundung zu minimalen Abweichungen (99,99% oder 100,01%), wurde auf exakt 100% gerundet.

Abbildung 5: Beteiligung LSW am Auffinden der Dokumente im deutschsprachigen Sample

Art der Beteiligung LSW	Anzahl gefundener Dokumente	%
Ausschließlich durch LSW	125	41,25
Durch LSW sowie mindestens eine weitere Variable	15	4,95
Nicht durch LSW	163	53,80
Gesamtsumme	303	100

Abbildung 6: Beteiligung LSW am Auffinden der Dokumente im fremdsprachigen Sample

Art der Beteiligung LSW	Anzahl gefundener Dokumente	%
Ausschließlich durch LSW	8	2,03
Durch LSW sowie mindestens eine weitere Variable	24	6,09
Nicht durch LSW	362	91,88
Gesamtsumme	394	100

Der Anteil der ausschließlich über eine einzige Variable gefundenen Dokumente an der Gesamtmenge der zu den jeweils 40 Anfragen pro Sample gefundenen Dokumente kann für das deutschsprachige Sample folgendermaßen hierarchisiert werden (siehe Abb. 7):

LSW:	125 Treffer (41,25%)
Kataloganreicherung:	71 Treffer (23,43%)
SW aus Fremddaten:	57 Treffer (18,81%)
Bibliographische Daten:	16 Treffer (5,28%)
Titel:	14 Treffer (4,62%)

Für das fremdsprachige Sample ergibt sich folgende Rangfolge (siehe Abb. 7):

Kataloganreicherung:	228 Treffer (57,87%)
Bibliographische Daten:	35 Treffer (8,88%)
Titel:	32 Treffer (8,12%)
SW aus Fremddaten:	16 Treffer (4,06%)
LSW:	8 Treffer (2,03%)

Abbildung 7: Ausschließlich durch eine einzige Variable gefundene Dokumente im Vergleich beider Samples

Variable	Anzahl gefundener Dokumente Sample deutsch-sprachig	%	Anzahl gefundener Dokumente Sample fremd-sprachig	%	Differenz in %
LSW	125	41,25	8	2,03	39,22
SW aus Fremddaten	57	18,81	16	4,06	14,75
Kataloganreicherung	71	23,43	228	57,87	-34,44
Weitere bibliographische Daten	16	5,28	35	8,88	-3,60
Titel	14	4,62	32	8,12	-3,50
Summe	283	93,40	319	80,96	12,44
Rest	20	6,60	75	19,04	-12,44
Gesamtsumme	303	100	394	100	

Bezogen auf die LSW ist der Anteil aufgefundener Dokumente damit im deutschsprachigen Sample rund 39% größer als im fremdsprachigen.
Die geringsten Unterschiede finden sich bei den Variablen bibliographische Daten und Titel, bei einer Differenz von 3,6% bzw. 3,5% zwischen den beiden Samples. Die zweitgrößte Abweichung lässt sich bei der Kataloganreicherung feststellen, deren Anteil am Auffinden der Dokumente im fremdsprachigen Sample gut 34% größer ist als im deutschsprachigen. Bei den SW aus Fremddaten ist im fremdsprachigen Sample demgegenüber ein um fast 15% geringerer Anteil an ausschließlich über diese Variable gefundenen Dokumenten zu beobachten als im deutschsprachigen.

Insgesamt wurde der allergrößte Teil der Treffer ausschließlich über eine einzige Variable generiert; im deutschsprachigen Sample in 93,4% der Fälle und im fremdsprachigen in 80,96%. Der verbleibende Anteil der Dokumente wurde durch mindestens eine weitere Variable aufgefunden.

Ergebnisse bei Berücksichtigung nur solcher Dokumente, die alle Variablen vorhalten

Betrachtet man den Anteil der ausschließlich über eine einzige Variable gefundenen Dokumente, die zudem über alle fünf Variablen verfügen, d.h. auch über SW aus Fremddaten und Kataloganreicherung, so verändert sich das Bild teilweise stark. Die Gesamtmenge der gefundenen Dokumente fällt unter dieser Einschränkung in allen drei Samples (gesamt, deutschsprachig, fremdsprachig) durchgehend niedriger aus. Bezogen auf

die Gesamtmenge aus beiden Samples kommt man nun zu folgender Hierarchisierung (siehe Abb. 8):

Kataloganreicherung: 223 Treffer (59,63%)
SW aus Fremddaten: 56 Treffer (14,97%)
LSW: 30 Treffer (8,02%)
Bibliographische Daten: 10 Treffer (2,67%)
Titel: 8 Treffer (2,14%)

Abbildung 8: Ausschließlich durch eine einzige Variable gefundene Dokumente in beiden Samples zusammengenommen bei Vorhandensein aller Variablen

Variable	Anzahl gefundener Dokumente	%
LSW	30	8,02
SW aus Fremddaten	56	14,97
Kataloganreicherung	223	59,63
Weitere bibliographische Daten	10	2,67
Titel	8	2,14
Summe	327	87,43
Rest	47	12,57
Gesamtsumme	374	100

Im deutschsprachigen Sample ergibt sich aus dieser Einschränkung folgende veränderte Rangfolge der Sucheinstiege (siehe Abb. 9):

Kataloganreicherung: 64 Treffer (42,38%)
SW aus Fremddaten: 44 Treffer (29,14%)
LSW: 30 Treffer (19,87%)
Bibliographische Daten: 4 Treffer (2,65%)
Titel: 4 Treffer (2,65%)

Auch beim fremdsprachigen Sample verändert sich die Hierarchie der Sucheinstiege bei Betrachtung der Teilmenge solcher Treffer, die alle Variablen vorhalten (siehe Abb. 9):

Kataloganreicherung: 159 Treffer (71,3%)
SW aus Fremddaten: 12 Treffer (5,38%)
Bibliographische Daten: 6 Treffer (2,69%)
Titel: 4 Treffer (1,79%)
LSW: 0 Treffer (0%)

Abbildung 9: Ausschließlich durch eine einzige Variable gefundene Dokumente im Vergleich beider Samples bei Vorhandensein aller Variablen

Variable	Anzahl gefundener Dokumente Sample deutsch-sprachig	%	Anzahl gefundener Dokumente Sample fremd-sprachig	%	Differenz in %
LSW	30	19,87	0	0,00	19,87
SW aus Fremddaten	44	29,14	12	5,38	23,76
Kataloganreicherung	64	42,38	159	71,30	-28,92
Weitere bibliographische Daten	4	2,65	6	2,69	-0,04
Titel	4	2,65	4	1,79	0,86
Summe	146	96,69	181	81,17	15,52
Rest	5	3,31	42	18,83	-15,52
Gesamtsumme	151	100	223	100	

Der Anteil der LSW am Auffinden der Dokumente ist im Abgleich der beiden Samples weiterhin deutlich geringer im fremdsprachigen Sample; allerdings nun nur noch um knapp 20% (gegenüber gut 39% Differenz beim Vergleich der Samples ohne das einschränkende Kriterium, dass alle Variablen vorhanden sind; siehe Abb. 7), was auf den großen Rückgang der durch LSW aufgefundenen Dokumente im deutschsprachigen Sample zurückgeführt werden kann. Die geringsten Unterschiede gibt es wieder bei den bibliographischen Daten und dem Titel, deren Differenz zwischen den Samples jeweils unter 1% liegt. Die größte Abweichung findet sich nun bei der Kataloganreicherung, die im fremdsprachigen Sample fast 29% mehr Treffer generiert als im deutschsprachigen. Anders verhält es sich mit den SW aus Fremddaten, die im deutschsprachigen Sample für fast 24% mehr Treffer verantwortlich sind als im fremdsprachigen.

Der Gesamtanteil der Treffer, die ausschließlich über eine einzige Variable gefunden wurden, bleibt weiterhin hoch und steigt noch ein wenig an; im deutschsprachigen Sample auf 96,69% der Fälle und im fremdsprachigen auf 81,17%.

Materialart

Im Hinblick auf die Materialart haben Monographien in allen Samples den größten Anteil. Bezogen auf die Gesamtmenge der Anfragen aus beiden Samples liegt er bei 73,46% (siehe Abb. 10). Beim fremdsprachigen Sample ist er mit 78,43% am höchsten und weicht um 11,43% vom Anteil im deutschsprachigen Sample ab (siehe Abb. 11). Unterschiede

zwischen den Samples finden sich v.a. im Hinblick auf analoge Zeitschriften, bei denen der Anteil im deutschsprachigen Sample rund 12% größer ist als im fremdsprachigen, sowie bei elektronischen Ressourcen, deren Anteil im fremdsprachigen Sample ca. 4% höher liegt als im deutschsprachigen (siehe Abb. 11). Eine gewisse Unschärfe bei der Auswertung der Daten zur Materialart ergibt sich aus dem Vorkommen solcher Dokumente mit Altsignaturen, da aus diesen Signaturen anders als bei denen nach Numerus Currens nicht hervorgeht, um welchen Medientyp es sich handelt.[153] Ihr Anteil ist jedoch äußerst gering, sodass eine mögliche Verzerrung der Daten als marginal angenommen werden kann.

Abbildung 10: Materialart der Dokumente in beiden Samples zusammengenommen

Materialart	Anzahl gefundener Dokumente	%
Monographien	512	73,46
Zeitschriften analog[154]	104	14,92
Zeitschriftenartikel analog	13	1,87
Altsignaturen	4	0,57
Elektronische Ressourcen	51	7,32
DVD	12	1,72
CD	1	0,14
Gesamtsumme	697	100

Abbildung 11: Materialart der Dokumente im Vergleich beider Samples

Materialart	Anzahl Dokumente Sample deutsch-sprachig	%	Anzahl Dokumente Sample fremd-sprachig	%	Differenz in %
Monographien	203	67,00	309	78,43	-11,43
Zeitschriften analog	66	21,78	38	9,64	12,14
Zeitschriftenartikel analog	9	2,97	4	1,02	1,95
Altsignaturen	4	1,32	0	0,00	1,32
Elektronische Ressourcen	15	4,95	36	9,14	-4,19
DVD	6	1,98	6	1,52	0,46
CD	0	0,00	1	0,25	-0,25
Gesamtsumme	303	100	394	100	

[153] Ausgehend von den Signaturen nach Numerus Currens sowie weitergehenden Angaben, die in den Excel-Tabellen in der linken Spalte (Suchterm) notiert wurden, etwa der PPN bei elektronischen Ressourcen oder dem Verweis „Zeitschriftenartikel“, konnte gezielt nach der Materialart gesucht werden, durch Filter wie „Z*/“, „A*/“, „PPN“ usw.

[154] Unter der Kategorie „Zeitschriften analog“ befinden sich vereinzelt auch Zeitungen, die mit einer „ZZ*/“-Signatur versehen sind. Eine Differenzierung dieser verschiedenen seriell erscheinenden Publikationen erschien für die hier vorgenommene Betrachtung nicht von Relevanz.

Dadurch, dass manche Materialarten nach 2016 weiter verschlagwortet werden, liegt die Vermutung nahe, dass der Anteil bestimmter Medientypen an den in den Testläufen aufgefundenen Dokumenten proportional höher liegt, als wenn nur solche Dokumente mit einer Zugangsnummer von vor 2016 berücksichtigt worden wären. Dies gilt, wie bereits in Kapitel 2.3. beschrieben, insbesondere für Zeitschriftentitel, DVDs und CDs sowie elektronische Ressourcen.[155] Es kann daher angenommen werden, dass gerade Suchanfragen mit großen Treffermengen viele der weiterhin verschlagworteten Medientypen unter den ersten 10 Trefferergebnissen vorhalten, da auch neuere Erscheinungen von 2016 und jüngeren Datums mit einer entsprechenden Zugangsnummer berücksichtigt wurden. Suchanfragen mit eher geringen Treffermengen weisen hingegen vermutlich öfter ältere Dokumente unter den ersten 10 Treffern nach, d.h. Dokumente die vor 2016 erschienen sind und auch vor 2016 erworben wurden.[156] Für die Frage nach der Bedeutung der LSW am Auffinden der verschiedenen Materialarten spielt diese mögliche Verzerrung jedoch keine weitere Rolle. Relevant könnte sie hingegen bei weiterführenden Analysen der Korrelationen zwischen den verschiedenen Sucheinstiegen werden, wie in Kapitel 4.3. weiter ausgeführt wird.[157]

[155] Einen Suchschlüssel zu finden, der bei der Suche im OPAC nach der Zugangsnummer filtert, stellte sich aufgrund verschiedener Sonderfälle und Ausnahmen als nicht sinnvoll heraus, da dann u.U. ganze Gruppen von Materialarten hätten ausgenommen werden müssen. Bei Zeitschriften etwa wird die Zugangsnummer dem Abonnement und nicht dem Titel zugewiesen, sodass sie für die einzelnen Zeitschriftentitel nicht recherchierbar ist.

[156] Ein Blick in die Dokumentation der Testläufe beider Samples scheint dies zu bestätigen (siehe Anhang 5); eine systematische Auswertung aller Suchanfragen ist an dieser Stelle jedoch nicht möglich. Ein weiterer Faktor, der Einfluss auf die Verteilung der Materialart auf die aufgefundenen Dokumente nehmen könnte, sind die in jüngerer Zeit verstärkt durchgeführten Digitalisierungsprojekte. Die Digitalisate werden als elektronische Ressourcen weiterhin verschlagwortet. Dadurch, dass in solchen Projekten in der Regel thematisch verwandte Originaldokumente digitalisiert werden, fällt auch die in diesen Fällen noch stattfindende Verschlagwortung meist sehr homogen aus und erhöht die Wahrscheinlichkeit gleich eine ganze Gruppe von Digitalisaten als Treffer zu einer thematisch passenden Suchanfrage aufzufinden. In den hier betrachteten Fällen bestätigt sich dies bezogen auf die zu den Digitalisierungsprojekten vergebenen LSW zwar nicht, dafür aber mit Blick auf die sehr ähnlichen bibliographischen Metadaten der entsprechenden Digitalisate; so geschehen bei den Suchanfragen „caricatura“ und „maestros“, die ausschließlich oder beinahe ausschließlich Dokumente aus einem Digitalisierungsprojekt zur spanischen Operettengattung der Zarzuela auffanden (siehe Anhang 9).

[157] Auch mit Blick auf die Indexierungsparameter ergeben sich Unterschiede, je nachdem welche Materialart vorliegt. So wird das Katalogfeld 046L bei DVDs mit ALL indexiert, für Textdokumente hingegen nicht. Auch diesem Aspekt eines Vergleichs der Daten mit und ohne Indexierungsparameter kann hier nicht weiter nachgegangen werden.

4. Analyse der Ergebnisse

Ausgehend von den im vorherigen Kapitel vorgestellten Ergebnissen lässt sich die Bedeutung der LSW beim Retrieval in einem mehrsprachigen Bestand differenzierter betrachten, wobei bei der Analyse auch weitergehende Aspekte aufgezeigt werden.

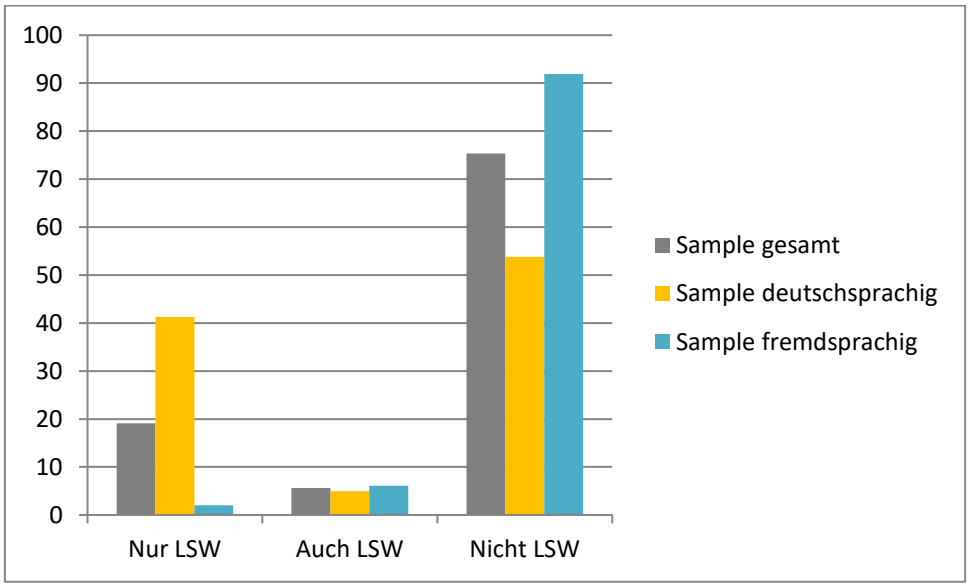
4.1. Bedeutung der lokalen Schlagworte

Bezogen auf die Gesamtmenge von 697 Dokumenten wurden 24,68% der Treffer mit Hilfe der LSW gefunden: 19,08% ausschließlich durch LSW, und 5,6% durch LSW und mindestens eine weitere Variable. In 75,32% der Fälle sind die Treffer ohne Beteiligung der LSW zu Stande gekommen (siehe Abb. 3).
D.h. knapp ein Fünftel der Dokumente (19,08%) wäre ohne die LSW gar nicht aufgefunden worden. Die LSW bilden damit den zweitwichtigsten Sucheinstieg nach der Kataloganreicherung (siehe Abb. 4). Bezogen auf die 80 Anfragen aus beiden Samples ergibt sich für die ausschließlich durch LSW gefundenen Dokumente ein Mittelwert von 1,66 Dokumenten pro Anfrage, wobei die Anzahl der pro Suchanfrage berücksichtigten Dokumente zwischen 1 und 10 variiert (siehe Anhang 12). Bei den Fällen, in denen neben den LSW auch weitere Variablen am Auffinden beteiligt waren, liegt der Mittelwert bei nur 0,49 Dokumenten pro Anfrage (siehe Anhang 12).

Ein deutlich differenzierteres Bild zeigt sich, wenn man die Samples für sich betrachtet. Je nachdem, in welcher Sprache die Suchanfragen gestellt wurden – auf Deutsch oder in einer Fremdsprache –, variiert der Anteil der jeweiligen Sucheinstiege im Hinblick auf die durch sie aufgefundenen Dokumente stark und ihre Hierarchie verändert sich.
So ist der Anteil der durch LSW aufgefundenen Dokumente im deutschsprachigen Sample deutlich größer. Hier wurden 46,2% der Dokumente unter Beteiligung der LSW gefunden, davon 41,25% ausschließlich durch diesen Sucheinstieg (siehe Abb. 5). Für das Sample deutschsprachiger Suchanfragen bedeutet dies, dass mehr als zwei Fünftel der Dokumente ohne LSW gar nicht gefunden worden wären. Im Vergleich aller Samples liegt hier der Mittelwert der ausschließlich durch LSW gefundenen Dokumente pro Suchanfrage mit 3,13 eindeutig am höchsten und weicht im Vergleich zu den Werten beider Samples zusammengenommen um 1,47 Punkte vom Mittelwert gefundener Dokumente pro Anfrage ab (siehe Anhang 13).

Deutlich anders verhält es sich mit der Bedeutung der LSW bei den fremdsprachigen Suchanfragen. Hier spielen die LSW kaum eine Rolle und sind insgesamt mit nur 8,12% am Auffinden der Dokumente beteiligt. Ausschließlich durch LSW gefunden wurden sogar nur 2,03%; d.h. nur eine sehr geringe Menge der Dokumente wäre ohne LSW gar nicht aufgefunden worden (siehe Abb. 6). Bezogen auf die Summe der Suchanfragen dieses Samples liegt der Mittelwert ausschließlich durch LSW gefundener Dokumente pro Anfrage bei lediglich 0,2 – er weicht damit um ganze 2,93 Punkte vom Mittelwert des deutschsprachigen Samples ab (siehe Anhang 14).
Die Beteiligung der LSW am Auffinden der Dokumente lässt sich im Vergleich aller Samples (gesamt: siehe Abb. 3, deutschsprachig: siehe Abb. 5, fremdsprachig: siehe Abb. 6) folgendermaßen graphisch darstellen, wobei die vertikale Achse den Anteil in Prozent angibt und die horizontale Achse den Grad der Beteiligung:

Abbildung 12: Beteiligung LSW am Auffinden der Dokumente in allen Samples

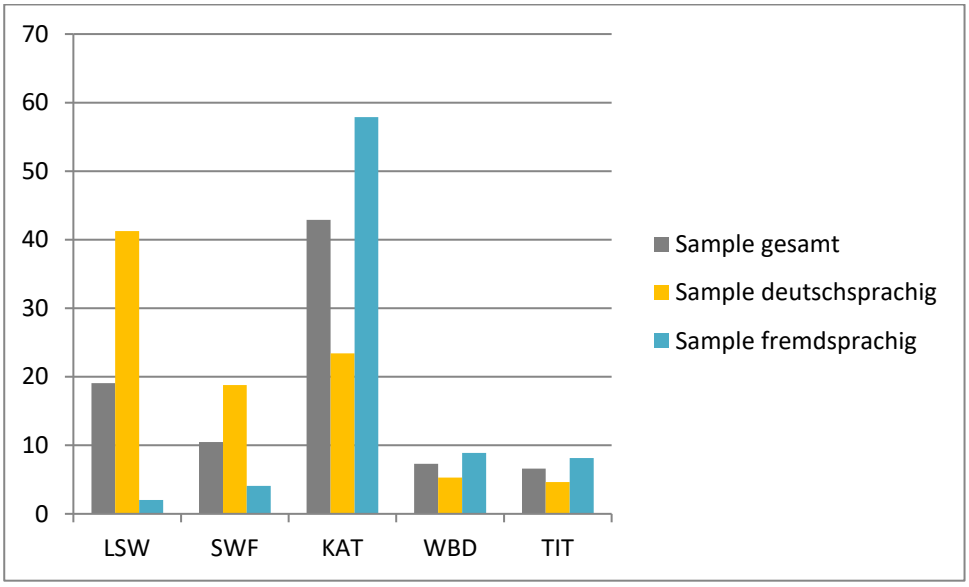


Schaut man sich den Anteil der ausschließlich über eine einzige Variable gefundenen Dokumente an und die damit verbundene Rangfolge der Sucheinstiege, so sind beim deutschsprachigen Sample die LSW mit über 40% der wichtigste Sucheinstieg, gefolgt von der Kataloganreicherung (siehe Abb. 7). Beim fremdsprachigen Sample hingegen ist die Kataloganreicherung mit fast 58% für die Mehrzahl der aufgefundenen Dokumente verantwortlich; die LSW nehmen hier mit einem Anteil von 2,03% allerdings die letzte Position noch hinter den SW aus Fremddaten ein (siehe Abb. 7). Die Differenz bei den Variablen LSW und Kataloganreicherung ist zwischen beiden Samples eindeutig am größten: Im deutschsprachigen Sample ist der Anteil der ausschließlich durch LSW

gefundenen Dokumente im Vergleich zum fremdsprachigen Sample um 39,22% größer, der Anteil der ausschließlich durch Kataloganreicherung gefundenen Dokumente dagegen um 34,44% kleiner. Es scheint folglich, dass gerade diese beiden Variablen – LSW und Kataloganreicherung – von Veränderungen bei der Eingabesprache maßgeblich betroffen sind.

Der Anteil der jeweils ausschließlich durch eine einzige der 5 Variablen aufgefundenen Dokumente lässt sich im Vergleich aller Samples (gesamt: siehe Abb. 4, deutschsprachig und fremdsprachig: siehe Abb. 7) folgendermaßen graphisch darstellen, wobei auf der vertikalen Achse die Angaben wieder in Prozent erfolgen:

Abbildung 13: Ausschließlich durch eine einzige Variable gefundene Dokumente in allen Samples



Insgesamt fällt auf, wie hoch der Gesamtanteil der ausschließlich über eine einzige Variable gefundenen Dokumente ist. Beim deutschsprachigen Sample wurden 93,4% der 303 Dokumente ausschließlich durch nur einen einzigen Sucheinstieg aufgefunden, beim fremdsprachigen Sample waren es 80,96% der 394 Dokumente (siehe Abb. 7).

Zwar wird die Relevanz der gefundenen Dokumente in dieser Arbeit nicht untersucht; allerdings kann vermutet werden, dass ein Retrievalergebnis eine größere Relevanz aufweist, wenn es über mehrere Sucheinstiege gleichzeitig gefunden wurde.[158] Ein

[158] Hier lohnt ein Blick in die durch die manuelle Auswertung ebenfalls verfügbaren Daten zu dem Anteil der fünf untersuchten Sucheinstiege am Auffinden der Dokumente ohne Berücksichtigung der von der VZG angewandten Indexierungsparameter: D.h. jedes Vorkommen eines Suchterms im Index wird gezählt, unabhängig davon, ob ein Exact Match besteht oder das entsprechende Indexfeld mit „ALL“ indexiert wurde. Unter diesen Bedingungen zeigt sich, dass im deutschsprachigen Sample nur noch 83,50% aller Dokumente ausschließlich durch eine einzige Variable gefunden wurden und im fremdsprachigen nur noch 75,13% (siehe Anhang 15). Das heißt im

Dokument, das sowohl durch LSW als auch SW aus Fremddaten gefunden wurde, spricht für eine konsistente intellektuelle Indexierung; ebenso legt eine Kombination aus Fundstellen in den Schlagworten, der Kataloganreicherung und dem Titel eine höhere thematische Relevanz eines Dokuments nahe, als wenn nur eines der Felder mit dem Suchbegriff indexiert wurde. Bei den bibliographischen Daten muss im Einzelfall genauer geschaut werden, inwieweit die Fundstellen inhaltsrelevant sind. So kann etwa bei der thematischen Suche nach Geographika das Auffinden des entsprechenden Suchterms als Verlagsort in der Veröffentlichungsangabe am Informationsbedürfnis der Nutzer_innen vorbeigehen. In den hier vorgenommenen Tests wurden bei Suchen nach Geographika – z.B. beim Suchterm „patagonien“ – Dokumente zudem teilweise ausschließlich durch den im Personennormsatz aus dem Formalerschließungsteil mitindexierten Geburtsort der Verfasser_innen aufgefunden (siehe Anhang 8 und 10), und man kann sich in solchen Fällen durchaus fragen, inwieweit diese Information dem Informationsbedürfnis der Nutzer_innen entspricht.[159]

Der Appell von Furnas et al. und anderen Autor_innen, möglichst viele Sucheinstiege zu schaffen, um Medien auf verschiedene Weisen auffindbar werden zu lassen, scheint also einerseits durchaus sinnvoll. Andererseits sollte eine solche Erhöhung der Sucheinstiege mit einer Gewichtung der Indexfelder verbunden sein, um eine Ordnung, Homogenisierung und Hierarchisierung der erzeugten Trefferergebnisse zu gewährleisten, die auch für die Nutzer_innen nachvollziehbar sein muss. Der OPAC des IAI bietet zwar die Möglichkeit einer Sortierung der Ergebnisse nach Relevanz, allerdings konnte im Rahmen dieser Arbeit keine Auswertung der Struktur des dahinterliegenden Algorithmus geleistet werden.[160] Des Weiteren zeigen die als Datenbasis der Suchanfragen dienenden Logdaten deutlich, dass eine Sortierung nach Relevanz, die eine bewusste Veränderung der Default-Einstellung der Suchmaske bedeutet, von den Nutzer_innen ohnehin kaum genutzt wird.[161]

Umkehrschluss, dass der Anteil der Dokumente, in denen ein Suchterm mehr als einmal indexiert wurde, wächst und damit u.U. auch die Relevanz, die solchen Dokumenten zugesprochen werden könnte.

[159] So im Falle der Signaturen A 15 / 25857 und A 17 / 16077. Es kann vermutet werden, dass die aufgefundenen Dokumente, die ansonsten keinen thematischen Bezug zu der argentinischen Region Patagonien aufweisen, nicht dem Informationsbedürfnis der Nutzer_innen entsprechen. Dies ist natürlich eine spekulative Annahme, die ausgehend von einer Einwortanfrage nicht abschließend beantwortet werden kann; eine Relevanzbewertung durch die Nutzer_innen wäre hier sicherlich aufschlussreich.

[160] Gewichtungsfaktoren bei der Relevanzsortierung sind nach Auskunft der VZG u.a. die Termfrequenz innerhalb einer Katalogaufnahme sowie die Länge des Unterfeldes, in dem sich der Term befindet.

[161] Aus den Daten des Logfiles blieben nach Ausschluss anderer Datenbanken als der des IAI sowie der internen IP-Adressen und bei Einschränkung auf Freitextsuchen, die nicht mit Zahlen beginnen, nur 135 Hits übrig, bei denen eine Sortierung nach Relevanz eingestellt wurde; siehe dazu Fußnote 139.

4.2. Bedeutung der lokalen Schlagworte in einem multilingualen Kontext

Aus dem großen Ungleichgewicht im Hinblick auf die Bedeutung der LSW für deutschsprachige gegenüber fremdsprachigen Anfragen lässt sich – unter dem Vorbehalt, dass weitergehende statistische Auswertungen ausstehen – vorsichtig schließen, dass die Anfragesprache – hier nur sehr grob unterschieden in deutsch- vs. fremdsprachig – eine Auswirkung darauf hat, wie die Trefferdokumente aufgefunden werden. Auch die in Kapitel 2.4. geäußerte Vermutung, dass die Sprache, in der die Suchanfragen jeweils verfasst sind, Auswirkungen auf die Bedeutung der LSW hat, scheint sich zu bestätigen.
Die in den Testläufen gewonnenen Ergebnisse zeigen deutlich, dass die LSW bei den deutschsprachigen Anfragen die meisten Treffer erzeugen. Bei den hier untersuchten fremdsprachigen, meist spanischsprachigen Suchanfragen spielen die überwiegend deutschsprachigen LSW hingegen kaum eine Rolle beim Auffinden der Dokumente.

Ein genauerer Blick auf die Suchanfragen, bei denen ein Suchterm in den LSW indexiert wurde, kann hier aufschlussreich sein. Im deutschsprachigen Sample sind 16 von den insgesamt 40 Anfragen (40%) für alle 125 ausschließlich durch LSW aufgefundenen Dokumente verantwortlich, wobei zu allen 16 Anfragen auch 10 Trefferdokumente ausgewertet werden konnten (siehe Anhang 13). In 3 der 16 Anfragen („afrobrasilianer“, „bergsteigen“, „nationalismus“) wurden alle 10 berücksichtigten Dokumente ausschließlich durch LSW gefunden; der studentisierte Wert beträgt hier 1,68 (siehe Abb. 14).
Bezogen auf die Gesamtmenge aller 80 Anfragen erhöht sich der studentisierte Wert bei diesen 3 Suchtermen auf 2,57 (siehe Anhang 12) – d.h. er weicht um fast eine Einheit vom studentisierten Wert des deutschsprachigen Samples ab. Daraus folgt, dass der Anteil solcher Fälle, in denen eine Suchanfrage die ersten 10 berücksichtigten Dokumente ausschließlich durch LSW aufgefunden hat, auf die Gesamtmenge aller 80 Anfragen betrachtet deutlich geringer ist.
Abbildung 14 zeigt diese 16 Suchanfragen aus dem deutschsprachigen Sample in aufsteigender Reihenfolge bezogen auf die Zahl der ausschließlich durch LSW aufgefundenen Dokumente. Die kursiv und fett gedruckt dargestellten Suchterme markieren, wenn Dokumente zugleich *ausschließlich durch LSW* sowie *durch LSW und mindestens eine weitere Variable* aufgefunden wurden.

Abbildung 14: Deutschsprachige Suchanfragen, die Dokumente ausschließlich durch LSW auffinden

Suchterm	Anzahl berücksichtigter Dokumente	Davon ausschließlich durch LSW gefunden	Studentisierter Wert
nachlass	10	2	-0,28
bolivien	10	5	0,95
kolumbien	10	6	0,95
patagonien	10	7	0,95
sicherheit	10	7	0,95
globalisierung	10	8	1,19
humor	10	8	1,19
religion	10	8	1,19
spionage	10	8	1,19
infrastruktur	10	9	1,44
katalonien	10	9	1,44
korruption	10	9	1,44
mediation	10	9	1,44
afrobrasilianer	10	10	1,68
bergsteigen	10	10	1,68
nationalsozialismus	10	10	1,68

Durch LSW sowie weitere Variablen gefundene Dokumente lassen sich in 7 von 40 Anfragen (17,5%) feststellen (siehe Anhang 13). Diese 7 Suchanfragen decken sich wie bereits ausgeführt mit einem Teil der 16 im vorherigen Abschnitt besprochenen Suchen, d.h. zu ihnen wurden zugleich Dokumente ausschließlich durch LSW aufgefunden. Der Maximalwert von 5 durch LSW und weitere Variablen gefundenen Dokumenten weicht mit einem studentisierten Wert von 4,4 deutlich vom Mittelwert ab, der nicht studentisiert bei 0,38 Dokumenten pro Anfrage liegt (siehe Anhang 13). Den Ausschnitt dieser 7 Suchterme aus der Gesamtmenge der Anfragen zeigt Abbildung 15, wobei die Suchterme wieder aufsteigend angeordnet sind, nach der Anzahl der durch LSW sowie weitere Variablen gefundenen Dokumente.

Abbildung 15: Deutschsprachige Suchanfragen, die Dokumente durch LSW sowie weitere Variablen auffinden

Suchterm	Anzahl berücksichtigter Dokumente	Davon durch LSW + weitere Variablen gefunden	Studentisierter Wert
globalisierung	10	1	0,59
korruption	10	1	0,59
religion	10	1	0,59
spionage	10	1	0,59
sicherheit	10	2	1,54
kolumbien	10	4	3,45
bolivien	10	5	4,40

Beim fremdsprachigen Sample zeigt sich ein deutlich anderes Bild. Dort finden sich lediglich in 5 von 40 Anfragen (12,5%) Dokumente, die ausschließlich durch LSW gefunden wurden (siehe Anhang 14), und ihr Anteil an der Gesamtmenge berücksichtigter Dokumente pro Anfrage ist geringer. Abbildung 16 zeigt diese 5 Suchterme, wieder in aufsteigender Anordnung:

Abbildung 16: Fremdsprachige Suchanfragen, die Dokumente ausschließlich durch LSW auffinden

Suchterm	Anzahl berücksichtigter Dokumente	Davon ausschließlich durch LSW gefunden	Studentisierter Wert
capoeira	10	1	1,43
mapuche	10	1	1,43
cuba	10	2	3,21
guarani	10	2	3,21
vanguardia	10	2	3,21

Betrachtet man die Fälle, in denen Dokumente durch LSW sowie weitere Variablen gefunden wurden, so lassen sich 6 von 40 Suchanfragen (15%) identifizieren (siehe Anhang 14), wobei auch hier alle Suchanfragen bis auf „nahuatl“ gleichzeitig Dokumente enthalten, die ausschließlich durch LSW gefunden wurden (siehe die kursivierten und fett gedruckten Suchterme in Abb. 17). Der Maximalwert liegt bei 8 durch LSW sowie weitere Variablen gefundenen Dokumenten in 2 der 6 Fälle. Der studentisierte Wert beträgt hier 3,98 und bedeutet damit eine deutliche Abweichung vom Mittelwert, der nicht studentisiert bei 0,6 Dokumenten pro Anfrage liegt (siehe Anhang 14). Abbildung 17 zeigt diesen Teilausschnitt an Suchanfragen, wobei die Suchterme wieder aufsteigend nach der Zahl aufgefundener Dokumente angeordnet sind:

Abbildung 17: Fremdsprachige Suchanfragen, die Dokumente durch LSW sowie weitere Variablen auffinden

Suchterm	Anzahl berücksichtigter Dokumente	Davon durch LSW + weitere Variablen gefunden	Studentisierter Wert
capoeira	10	1	0,22
nahuatl	10	1	0,22
vanguardia	10	2	0,75
guarani	10	4	1,83
cuba	10	8	3,98
mapuche	10	8	3,98

Insgesamt fällt auf, dass die LSW – aber auch solche aus Fremddaten – einen großen Anteil am Retrieval nehmen, sobald sie deckungsgleich mit den Suchtermen sind. Für das deutschsprachige Sample wurde bereits auf 3 Suchanfragen verwiesen, bei denen alle berücksichtigten Dokumente ausschließlich durch ein entsprechendes LSW gefunden wurden. Weniger absolut, aber dennoch in sehr großem Maße am Auffinden der Dokumente beteiligt sind LSW in 12 der insgesamt 16 Suchanfragen, in denen ein Teil der Dokumente entweder ausschließlich durch LSW oder durch LSW und weitere Variablen gefunden wurde. Dies gilt für die Suchterme „bolivien“, „globalisierung“, „humor“, „infrastruktur“, „katalonien“, „kolumbien“, „korruption“, „mediation“, „patagonien“, „religion“, „sicherheit“ und „spionage“. Bei diesen Suchtermen wurde die Mehrheit aller Dokumente, d.h. mehr als 5 Dokumente, entweder ausschließlich durch LSW oder im Zusammenspiel mit weiteren Indexfeldern aufgefunden. Lediglich der Suchterm „nachlass“ bildet in dieser Hinsicht eine Ausnahme, da hier 8 von 10 Dokumenten ohne jegliche Beteiligung der LSW gefunden wurden.
Bei den fremdsprachigen Suchtermen hingegen findet sich keine Suchanfrage, bei der alle Dokumente ausschließlich durch LSW gefunden wurden. Überdurchschnittliche Ergebnisse weisen die Suchanfragen „cuba“, „guarani“ und „mapuche“ auf, bei denen mehr als die Hälfte der Dokumente entweder ausschließlich durch LSW oder durch LSW sowie weitere Variablen aufgefunden wurde. Für die Suchanfragen „capoeira“ und „vanguardia“ hingegen wurden nur 2 bzw. 4 Dokumente unter Beteiligung von LSW gefunden.
Bei den Suchanfragen „guarani“ und „mapuche“ könnte ein Faktor für den großen Anteil der durch LSW aufgefundenen Dokumente sein, dass zu diesen Themen eine große Zahl an Schlagwortvarianten vorliegt. Die semantische Auffächerung in verschiedene Schlagworte leistet eine thematische Differenzierung, je nachdem, ob mit der Bezeichnung die Sprachgruppe, Texte in den entsprechenden Sprachen oder die Ethnien, die diese Sprachen sprechen, gemeint sind. Durch eine derartige Disambiguierung der ansonsten ambiguen Bezeichnungen können sehr viele Dokumente in differenzierter Weise dem Themenkomplex „guarani“ bzw. „mapuche“ zugeordnet werden. Die Dokumente werden dann durch ein entsprechendes Schlagwort auffindbar – auch dann, wenn es sich z.B. um eine sehr spezifische linguistische Auseinandersetzung mit nur einer der vielen Mapuche-Sprachen handelt und der Term „mapuche“ in den bibliographischen Metadaten oder den ToC nicht indexiert ist.

Ähnliches ist bei den SW aus Fremddaten zu beobachten, die hier zwar nicht im Fokus stehen, für die Ermittlung des Anteils der LSW am Retrieval jedoch auch dokumentiert wurden. Im deutschsprachigen Sample etwa wurden alle Dokumente zu den Suchanfragen „femizid“, „sprachmittlung“ und „subkultur“ ausschließlich durch SW aus Fremddaten gefunden, wobei der Suchterm „femizid“ anders als die anderen beiden Beispiele insgesamt

nur 2 und nicht 10 Treffer oder mehr erzeugte. Des Weiteren finden sich auch bei den SW aus Fremddaten Fälle, in denen die kontrollierten Vokabulare die Mehrheit der Treffer erzeugten, z.B. bei den Suchtermen „auslandsschule“ oder „gegenkultur“ mit 6 von 7 bzw. 7 von 10 berücksichtigten Dokumenten. Beim Suchterm „photo“ wurde genau die Hälfte der 10 berücksichtigten Dokumente ausschließlich durch SW aus Fremddaten gefunden.
Im fremdsprachigen Sample kann hierfür lediglich das Beispiel des Suchterms „soccer“ herangezogen werden, der in 7 von 10 berücksichtigten Dokumenten ausschließlich durch SW aus Fremddaten gefunden wurde und in 2 Dokumenten durch SW aus Fremddaten sowie weitere Variablen. Dieses letzte Einzelbeispiel zeigt darüber hinaus, wie wichtig das Matching zwischen Suchsprache und Indexierungssprache (in diesem Fall eines kontrollierten Vokabulars) ist. Ohne die inhaltliche Erschließung durch ein in diesem Fall englischsprachiges Schlagwort wäre keines der 7 ausschließlich durch SW aus Fremddaten gefundenen Dokumente aufgefunden worden, da sie alle auf Spanisch oder Portugiesisch verfasst sind und den Suchterm weder in den bibliographischen Metadaten noch im ToC enthalten (die Sprache der Dokumente kann den Screenshots aus der Dokumentation zu dem Suchterm „soccer“ in Anhang 5 entnommen werden). Das einzige englischsprachige Dokument unter den 10 berücksichtigten Treffern wurde im Gegensatz dazu durch die Indexierung des ToC gefunden, da in diesem Fall Suchsprache und Objektsprache übereinstimmen.

Schaut man sich die im fremdsprachigen Sample insgesamt 32 Dokumente (siehe Abb. 6), die ausschließlich oder unter anderem durch LSW gefunden wurden, daraufhin an, welche LSW hier zum Tragen kamen, so fällt auf, dass es sich in fast allen Fällen um Eigenbezeichnungen handelt. Insgesamt waren im fremdsprachigen Sample 12 LSW am Auffinden von Dokumenten beteiligt (siehe Anhang 11); davon sind 11 entweder Geographika (Cuba), Bezeichnungen von Musik- oder Tanzstilen (Capoeira) oder Variationen von Bezeichnungen sprachlicher oder ethnischer Zugehörigkeiten (Guaraní, Náhuatl und Mapuche). Einzige Ausnahme ist das LSW „Avantgarde“, das in seinem Normsatz die spanische Übersetzung „Vanguardia“ enthält, die sich mit dem eingegebenen Suchterm deckt.
In zwei weiteren Fällen sind ebenfalls spanische Übersetzungen in den Normsätzen der LSW hinterlegt, die sich thematisch mit den hier getesteten Suchtermen decken: „favelas“ innerhalb des LSW „Elendsviertel“ und „galegos“ im LSW „Galicier“ (siehe Anhang 11). Allerdings kommen hier die Begrenzungen des auf dem Exact-Match-Retrieval aufbauenden OPAC-Systems zum Tragen, das diese thematisch relevanten Terme nicht mit den jeweils im Singular gestellten Anfragen matchen kann (die Anfragen lauteten „favela“ und galego“).

Diese Einzelbeispiele zeigen ebenso das Potential multilingualer Erweiterungen von Thesauri wie auch die durch das Exact Match erzeugten Hindernisse, die es erforderlich machen, indexierte Begriffe jeweils in Singular- und Pluralform vorzuhalten.

In den SW aus Fremddaten, die sehr häufig aus der GND übernommen werden, finden sich Beispiele für eine solche doppelte Indexierung mit den sich dadurch ergebenden Vorteilen für das Retrieval. So wird die Singularform „favela“ im Normsatz des GND-Sachbegriffs „Slum“ gefunden, in dem sowohl Singular- als auch Pluralform aufgeführt sind.[162]

Die Anforderung einer exakten Übereinstimmung gilt natürlich auch für die deutschsprachigen Suchanfragen und lässt sich am Beispiel des Suchterms „photo“ nachvollziehen. Während im Normsatz des LSW „Photographie“ neben den Schreibvarianten mit „f“ nur das Synonym „Photos“ im Plural enthalten ist, führt der Normsatz des GND-Sachbegriffs „Fotografie“ die Schreibvariante „Photo“ auf, dafür aber nicht die Pluralform. Trotz der thematischen Nähe der hier referierten Schlagworte ist damit die – so ist anzunehmen – häufig willkürliche Form der Eingabe als Singular oder Plural entscheidend im Hinblick darauf, ob ein Schlagwort einen Treffer zu einer Anfrage erzeugen kann.

Gleiches gilt für Schreibweisen, bei denen der Suchterm in einer längeren Zeichenkette enthalten ist, die – dem Prinzip des Exact Match folgend – in der exakt vorliegenden Form indexiert wird. So wurde der Suchterm „religion“ beispielsweise nicht durch das LSW „Gewissensfreiheit <Recht>“ aufgefunden (Z / 539 : 108(2017)), in dessen Normsatz auch das Wort „Religionsfreiheit“ enthalten ist; ebenso wenig der Suchterm „spionage“ in englischsprachigen Schlagwortketten wie „Espionage -- Mexico -- History -- 20th century“ (Signatur A 15 / 21026) (siehe dazu Anhang 10).

Die in der Forschungsliteratur beschriebenen Probleme der Nutzer_innen im Umgang mit OPACs können auch für den Online-Katalog des IAI als relevant angenommen werden, da die wenigsten Nutzer_innen mit Suchhilfen wie Boole´schen Operatoren oder Trunkierung vertraut sind, ganz zu schweigen von den bei der Indexierung angewandten Suchschlüsseln, die das gezielte Durchsuchen von Katalogfeldern erlauben, die u.U. nicht mit „Alle Wörter [ALL]“ indexiert sind. Ein für den multilingualen Kontext des IAI relevanter Fall könnte das Anmerkungsfeld 046L sein, das häufig dazu genutzt wird, Besonderheiten im Hinblick auf die Sprache des vorliegenden Dokuments zu erfassen, und das für Monographien nicht mit „ALL“ indexiert wird.[163] Gerade derartige Informationen könnten jedoch mit Blick auf das

[162] Die hier genannten GND-Sachbegriffe können in der Online-Präsenz der GND, der OGND, eingesehen werden; siehe: http://swb.bsz-bw.de/DB=2.104/ (zuletzt geprüft am 11.03.2020).

[163] Dies ist umso relevanter, da auch das Feld „Sprache/n“ aus der OPAC-Anzeige nicht nach den tatsächlichen Sprachbezeichnungen durchsuchbar ist, da die Sprachangaben ausgehend von Codes, die in dem entsprechenden Katalogfeld im LBS erfasst werden, automatisch generiert werden. Zwar gibt es die Möglichkeit in der erweiterten Suche nach Sprachen zu filtern; inwieweit diese Option von den Nutzer_innen auch genutzt wird, kann ich ohne weitere Erhebungen allerdings nicht beurteilen.

Informationsbedürfnis multilingualer Nutzer_innengruppen oder auch für die Beantwortung spezifischer Forschungsfragen zentral sein, z.B. wenn nach Lyrik in einer bestimmten indigenen Sprache gesucht wird.

Das Problem solcher Schreibvarianten verstärkt sich naturgemäß, sobald Sprachgrenzen überschritten werden. Das Potential, das sich für die LSW aus einer multilingualen Erweiterung ergeben könnte, muss folglich als sehr groß angenommen werden. Insbesondere bei den Geographika könnte eine Homogenisierung der Suchergebnisse unabhängig von der Anfragesprache erzielt werden, indem die Normsätze der LSW systematisch immer auch die Bezeichnung des jeweiligen Landes in seiner eigenen Sprache (d.h. zumeist Spanisch oder Portugiesisch) enthalten. Dies ist jedoch in den seltensten Fällen gegeben – etwa bei Kuba/Cuba. Anders verhält es sich beispielsweise bei dem Begriffspaar Bolivien/Bolivia, das sowohl im fremdsprachigen Sample mit dem Suchterm „bolivia“ getestet wurde als auch im deutschsprachigen unter der Bezeichnung „bolivien“. Die Ergebnisse beider Anfragen waren dementsprechend unterschiedlich und nur in 3 Fällen deckungsgleich (Signaturen Z / 30295, Z / 29863, Z / 29853). In allen drei Fällen haben wir es mit fremdsprachigen Dokumenten auf Spanisch zu tun, die in ihren Metadaten – im Titel oder der Veröffentlichungsangabe – die Bezeichnung „Bolivia“ enthalten, für das deutschsprachige Retrieval jedoch einzig und allein durch das deutschsprachige LSW „Bolivien“ als thematisch zugehörig markiert und auffindbar gemacht wurden. Eine Zusammenführung der ansonsten voneinander getrennten Bezeichnungen und die daraus folgende Homogenisierung der Trefferergebnisse unabhängig von der eingesetzten Suchsprache scheinen insofern sinnvoll und aus Sicht der – häufig mehrsprachigen – Forschenden wünschenswert.

Eine Zusammenführung thematisch verwandter Deskriptoren stellt sich als deutlich schwieriger heraus, da hier häufig keine Eins-zu-eins-Relationen zwischen den Begriffen und Bezeichnungen der verschiedenen Sprachen bestehen (siehe dazu Kapitel 2.2.). Ein Mapping von Vokabularen in einem nicht-symmetrischen Thesaurus wäre aber durchaus eine Möglichkeit, Deskriptoren über Sprachgrenzen hinweg thematisch zusammenzuführen.

Weitere thematisch zusammengehörige Suchanfragen innerhalb der getesteten Samples sind „femizid/femicide“, „kolumbien/colombia“, „nationalsozialismus/nacionalsocialismo“ und „quietismus/quietismo“ – das erste Wort dieser Entsprechungen entstammt dabei dem deutschsprachigen Sample, während das zweite fremdsprachig ist, im ersten Fall potentiell englischsprachig und in den anderen drei Fällen potentiell aus dem Spanischen oder Portugiesischen. Schaut man sich in diesen Fällen die Objektsprache der aufgefundenen Dokumente an (siehe hierzu die Screenshots aus der Dokumentation zu den entsprechenden Anfragen in Anhang 5), so kommt man zu folgendem Ergebnis:

femizid: 2 von 2 berücksichtigten Dokumenten sind fremdsprachig (Spanisch, Englisch)

femicide: 10 von 10 berücksichtigten Dokumenten sind fremdsprachig (Englisch, Spanisch, Französisch, Italienisch)

kolumbien: 3 von 10 berücksichtigten Dokumenten sind deutschsprachig, 7 Dokumente sind fremdsprachig (Spanisch)

colombia: 10 von 10 berücksichtigten Dokumenten sind fremdsprachig (Spanisch)

nationalsozialismus: 10 von 10 berücksichtigten Dokumenten sind fremdsprachig (Spanisch, Englisch, Portugiesisch)

nacionalsocialismo: 10 von 10 berücksichtigten Dokumenten sind fremdsprachig (Spanisch, Portugiesisch)

quietismus: 2 von 3 berücksichtigten Dokumenten sind deutschsprachig, 1 Dokument ist fremdsprachig (Spanisch)

quietismo: 4 von 4 berücksichtigten Dokumenten sind fremdsprachig (Spanisch)

Wenig überraschend ist die Tatsache, dass die fremdsprachigen Suchterme in keinem einzigen der hier untersuchten Fälle deutschsprachige Dokumente unter den berücksichtigten Treffern aufweisen. Abgesehen von den wenigen Ausnahmen, in denen die Schlagwortnormsätze Übersetzungen in andere Sprachen enthalten oder das Schlagwort eine ursprünglich fremdsprachige Eigenbezeichnung ist – wie bereits weiter oben an anderen Beispielen ausgeführt wurde – können die überwiegend deutschsprachigen Schlagworte nicht auf fremdsprachige Suchanfragen abgebildet werden. Dies bestätigt sich mit Blick auf die 4 hier betrachteten Suchanfragen aus dem fremdsprachigen Sample, zu denen kein einziges Dokument über LSW gefunden wurde.

Bei den deutschsprachigen Anfragen hingegen wird sichtbar, dass die intellektuelle Erschließung eine thematische Zuordnung unabhängig von der Objektsprache ermöglicht. So finden sich hier gleichsam deutschsprachige und fremdsprachige Dokumente unter den Ergebnissen. Allerdings muss von Seiten der Nutzer_innen die Sprache der Suchanfragen deckungsgleich mit der des kontrollierten Vokabulars sein, um die durch die Schlagworte thematisch gebündelten Dokumente auch auffinden zu können. Bei den deutschsprachigen Anfragen ist dies beim Suchterm „kolumbien“ gelungen, in dessen Fall alle 10 Dokumente ausschließlich durch oder unter Beteiligung von LSW gefunden wurden, sowie für den

Suchterm „nationalsozialismus“, in dessen Fall sogar alle 10 Dokumente ausschließlich durch LSW gefunden wurden.

Ausgehend von diesen Einzelfällen zeigt sich, wie aufschlussreich eine systematische Auswertung der Dreiecksbeziehung Eingabesprache/Dokumentationssprache/Objektsprache sein könnte. Die wenigen Einzelbeispiele, die hier genannt wurden, legen den Schluss nahe, dass durch ein solches Vorgehen noch weit differenziertere Aussagen darüber getroffen werden könnten, in welchen Fällen und für welche Nutzer_innen die LSW – oder auch andere Variablen (insbesondere die Kataloganreicherung) – von Nutzen sind.[164]

[164] Aus den hier gesammelten Daten könnten alle diesbezüglich relevanten Angaben erhoben werden, indem die Screenshots der berücksichtigten Dokumente, die ausschließlich oder auch durch LSW gefunden wurden, auf die Sprache geprüft würden, in der die Dokumente verfasst sind. Für die anderen Variablen könnte analog vorgegangen werden. Dies würde jedoch die hier verfolgte Forschungsfrage sprengen und einen weiteren, sehr aufwändigen Schritt in der Dokumentation bedeuten, der in der Kürze der Bearbeitungszeit nicht zu leisten gewesen wäre.

4.3. Lokale Schlagworte und Kataloganreicherung

Ebenfalls von Interesse für eine Bewertung der Bedeutung der LSW beim Retrieval scheint der Blick auf die Teilmenge solcher Dokumente, die ausschließlich über einen einzigen Sucheinstieg aufgefunden wurden und bei denen alle 5 Variablen vorliegen, d.h. auch SW aus Fremddaten und Kataloganreicherung. In beiden Samples zusammengenommen werden in dieser Teilmenge von 374 der insgesamt 697 Dokumente nur noch 8,02% der Treffer ausschließlich durch LSW gefunden (siehe Abb. 8).

Bei der Kataloganreicherung und den SW aus Fremddaten, die anders als die anderen Variablen unter Umständen fehlen können, zeigt sich eine gegenläufige Tendenz; sind alle Variablen vorhanden, so steigt der Anteil der ausschließlich durch sie gefundenen Treffer: bei der Kataloganreicherung von 42,9% auf 59,63% und bei den SW aus Fremddaten von 10,47% auf 14,97% (siehe Abb. 4 und Abb. 8). Ein Anstieg ist insofern erwartbar, da diese beiden Variablen auch nur bei ihrem Vorhandensein die Möglichkeit haben, Dokumente aufzufinden. Für die SW aus Fremddaten hat dieser Anstieg zur Folge, dass ihre Bedeutung gegenüber den LSW steigt und sie nun den zweitwichtigsten Sucheinstieg bilden (siehe Abb. 8).

Dieses zusätzliche Kriterium hat in beiden Samples Veränderungen in der Rangfolge der Sucheinstiege zur Folge. Im deutschsprachigen Sample senkt es den Anteil der ausschließlich durch LSW gefundenen Dokumente von vormals 41,25% – bezogen auf die Gesamtmenge von 303 Dokumenten (siehe Abb. 7) – auf nur noch 19,87% – bezogen auf die Teilmenge von 151 Dokumenten (siehe Abb. 9). Dies bedeutet einen Rückgang um 21,38% und lässt die LSW vom wichtigsten Sucheinstieg beim deutschsprachigen Retrieval auf die dritte Position hinter die Kataloganreicherung und die SW aus Fremddaten fallen. Diese beiden Variablen hingegen erleben, bezogen auf die Teilmenge der 151 Dokumente, die alle Variablen vorhalten, einen Anstieg gegenüber der Gesamtmenge von 303 Dokumenten: bei den SW aus Fremddaten um 10,33% und bei der Kataloganreicherung um ganze 18,95% (vgl. hierzu Abb. 7 und Abb. 9).

Festzuhalten bleibt aber, dass im deutschsprachigen Sample trotz des wesentlichen Rückgangs des Anteils der ausschließlich durch LSW gefundenen Dokumente noch immer knapp ein Fünftel der Dokumente ohne LSW gar nicht aufgefunden worden wären.

Mit Blick auf die 40 deutschsprachigen Anfragen verringert sich der Anteil solcher Suchterme, die Dokumente ausschließlich durch LSW oder durch LSW und weitere Variablen auffinden, von vormals 16 auf nun 11 Anfragen. Die Suchterme „bolivien“, „globalisierung“, „katalonien“, „kolumbien“ und „religion“ erzielen unter Ausschluss von Dokumenten ohne Kataloganreicherung und SW aus Fremddaten keinerlei Treffer mehr

unter Beteiligung von LSW und auch bei den anderen Suchtermen nimmt die Menge der durch sie aufgefundenen Dokumente ab (diese Angaben können Anhang 8 entnommen werden).

Beim fremdsprachigen Sample verändert sich die Hierarchie der Sucheinstiege ebenfalls. Die LSW bleiben auf der letzten Position; der Anteil der ausschließlich durch LSW aufgefundenen Dokumente sinkt aber von 2,03% bezogen auf die Gesamtmenge der 394 Dokumente (siehe Abb. 7) auf 0% bezogen auf die Teilmenge von 223 Dokumenten (siehe Abb. 9). Die SW aus Fremddaten wechseln hingegen von der vierten auf die zweite Position in der Rangordnung. Bei der Kataloganreicherung als durchgängig wichtigstem Sucheinstieg bei fremdsprachigen Suchen ergibt sich ein weiterer Anstieg des prozentualen Anteils um 13,43% gegenüber ihrem Anteil bezogen auf die Gesamtmenge von 394 Dokumenten (vgl. hierzu Abb. 7 und Abb. 9).
In diesem Szenario wären im fremdsprachigen Sample folglich alle Dokumente auch ohne die LSW gefunden worden. Die eher geringe Bedeutung dieses Sucheinstiegs im fremdsprachigen Sample wird so noch einmal untermauert und verschärft. Schaut man auf die betroffenen Suchanfragen, so zeigt sich, dass von vormals 6 Suchtermen, die bei den Trefferdokumenten ausschließlich oder auch in den LSW indexiert wurden, zwei Suchterme gänzlich wegfallen. So generieren die Suchterme „capoeira“ und „cuba“ bei Ausschluss solcher Dokumente ohne Kataloganreicherung oder SW aus Fremddaten keinerlei Treffer mehr durch LSW (diese Angaben können Anhang 9 entnommen werden).

Festgehalten werden muss also, dass bei Vorhandensein aller Variablen die Bedeutung der LSW am Retrieval deutlich absinkt. Aufgrund des deutlichen Anstiegs an Treffern durch Kataloganreicherung in diesem Szenario könnte man vermuten, dass v.a. dieser Sucheinstieg für die veränderten Verhältnisse verantwortlich ist. Hierbei gilt es jedoch wieder zwischen beiden Samples zu differenzieren. Beim deutschsprachigen Sample findet sich, bezogen auf die Teilmenge der 151 Dokumente, die alle Variablen vorhalten, nur ein einziges Dokument, das neben den LSW noch eine weitere Variable bedient, und zwar SW aus Fremddaten (Signatur A 14 / 9139, gefunden durch den Suchterm „spionage“, siehe Anhang 8).[165] Die verbleibenden 30 Dokumente werden ohnehin ausschließlich durch LSW aufgefunden. Des Weiteren kann (wie in Kapitel 4.2. ausgeführt) vermutet werden, dass ein großer Teil der zu den deutschsprachigen Suchanfragen gefundenen Dokumente nicht auf

[165] Die Excel-Tabelle aus Anhang 8 ermöglicht durch gezieltes Filtern die hier angeführten Ergebnisse zu reproduzieren. Durch entsprechende Filter können die mit „1“ markierten roten Felder bei der Kataloganreicherung und den SW aus Fremddaten ausgeschlossen werden und für die LSW nur die grünen Felder, die eine „1“ enthalten, angezeigt werden.

Deutsch verfasst ist und daher wahrscheinlich in den ToC die deutschsprachigen Suchterme nicht vorhält; eine systematische Untersuchung der Dimension der Objektsprache steht wie bereits erwähnt allerdings aus.

Beim fremdsprachigen Sample hingegen enthalten bei Vorhandensein aller Variablen 7 Dokumente den Suchterm in den LSW, wobei alle mindestens einen, meist jedoch gleich mehrere Fundstellen in den Indexfeldern der anderen Sucheinstiege aufweisen.[166]

Ein direkter Zusammenhang zwischen dem Rückgang der ausschließlich durch LSW gefundenen Dokumente und dem Anstieg bei der Kataloganreicherung bestätigt sich für die hier vorliegenden Daten folglich nicht.

Es kann in der Regel davon ausgegangen werden, dass Dokumente mit ToC eine Kataloganreicherung erfahren haben und nur solche ohne ToC diese Variable nicht bedienen – mit Ausnahme älterer Publikationen von vor 2000, die insgesamt jedoch in den Trefferdokumenten eher selten vertreten sind. Filtert man hier die Ergebnisse für das deutschsprachige Sample genauer, so zeigt sich, dass von den 125 Dokumenten, die aus der Gesamtmenge von 303 Dokumenten ausschließlich durch LSW gefunden wurden, 61 nicht über Kataloganreicherung verfügen; von diesen 61 Dokumenten wiederum sind 52 Zeitschriftentitel, bei denen nur in Ausnahmefällen, etwa bei Stücktiteln, das ToC gescannt wird, oder um Zeitschriftenartikel, die grundsätzlich kein ToC haben.[167] Es scheint also durchaus Korrelationen mit der Materialart der durch Kataloganreicherung gefundenen Dokumente zu geben. Diesem Aspekt kann hier allerdings nicht weiter nachgegangen werden, etwa durch eine Ermittlung des Anteils an Tonträgern oder Filmmaterial, die in der Regel ebenfalls keine ToC vorweisen. Da gerade diese Materialien auch nach 2016 weiter verschlagwortet werden und normalerweise nicht über Kataloganreicherung verfügen, kann vermutet werden, dass die LSW beim Auffinden der Dokumente insbesondere für diese Materialarten eine größere Rolle spielen.

Der große Anteil an Dokumenten, die durch Kataloganreicherung aufgefunden wurden, macht diese Variable in beiden Samples zusammengenommen und insbesondere im fremdsprachigen Sample eindeutig zum wichtigsten Sucheinstieg. Allerdings sind mit diesen Häufigkeitsangaben keinerlei Aussagen zur Relevanz der Treffer verbunden.

---

[166] Um zu diesem Ergebnis zu gelangen, wurde in der Excel-Tabelle aus Anhang 9 analog zu dem in Fußnote 165 beschriebenen Verfahren vorgegangen.

[167] Hierfür wurde in der Excel-Tabelle aus Anhang 8 in der Spalte „LSW“ nach den grünen Feldern, die „1“ enthalten, gefiltert, während bei allen verbleibenden Variablen diese Felder ausgeschlossen wurden; bei der Kataloganreicherung wurde außerdem nach den roten Felder, die „1“ enthalten, gefiltert. In der Spalte „Suchbegriff“ wurde nach „Z*/“ gefiltert, wodurch nur Signaturen von analogen Zeitschriftentiteln und –artikeln angezeigt werden.

Zwar kann bei den gescannten ToC, bei denen die Indexterme direkt aus dem Volltext extrahiert werden, ein inhaltlicher Zusammenhang angenommen werden; die kontextuelle Relevanz wird dabei jedoch nicht erfasst, ebenso wenig Fälle von Polysemie oder anderen Formen sprachlicher Ambiguität. So kann der Suchterm in einem sehr kurzen Unterkapitel gefunden werden oder in äußerst spezifischen Kontexten, die am Informationsbedürfnis der Nutzer_innen vorbeigehen können. Der Treffer ist dann zwar in einem weiteren Sinne thematisch passend, für den jeweiligen Nutzer/die jeweilige Nutzerin u.U. jedoch dennoch nicht relevant. Problematischer sind Fundstellen in den ToC, die auf die sprachliche Ambiguität von Bezeichnungen zurückgehen und bei der Extraktion von Termen durch Volltextindexierung anders als in kontrollierten Vokabularen nicht disambiguiert werden. Ein plastisches Beispiel hierfür bietet die Suchanfrage „barroco“ (spanisch und portugiesisch für Barock), die u.a. einen Treffer erzielt, der ausschließlich über die Kataloganreicherung gefunden wurde (Signatur A 13 / 22383 : 3) (siehe Anhang 9); allerdings hat das aufgefundene Dokument keinerlei Bezug zu der Epoche des Barock, sondern befasst sich mit der Beziehung von Mensch und Umwelt im Amazonasgebiet. „Barroco“ wird hier gefunden, da der Autor eines Unterkapitels diesen Namen trägt. Das Informationsbedürfnis der Nutzer_innen und in der Folge die Relevanz des aufgefundenen Treffers lassen sich an dieser Stelle nicht abschließend klären; das Problem der sprachlichen Ambiguität sollte aber dennoch deutlich geworden sein.[168]

Der sehr hohe Anteil der Kataloganreicherung am Gesamtergebnis würde sich bei einer Rückkopplung an eine Relevanzbewertung eventuell relativieren, denn Dokumente, die einen Themenbereich nur sehr marginal abdecken oder die nur aufgrund sprachlicher Ambiguität eine bestimmte Bezeichnung aus diesem Bereich vorhalten, würden in der intellektuellen Indexierung sehr wahrscheinlich nicht diesem Themengebiet zugeordnet werden. D.h. die Frage, der nachzugehen wäre, ist, ob die entsprechenden Dokumente nicht mit Schlagworten versehen wurden, da bei der intellektuellen Sacherschließung, die personengebunden und damit zugleich sehr individuell ist, ein thematischer Zusammenhang übersehen wurde oder da dieser Zusammenhang auf die Gesamtheit des Dokuments hin betrachtet zu vernachlässigen ist.

Eine differenzierte Erschließung von Dokumenten auch auf der Ebene von Unterabschnitten bietet zweifelsohne große Vorzüge. Problematisch ist jedoch das Nebeneinander der verschiedenen Indexierungsformen ohne jegliche Form der Gewichtung. Eine Gewichtung

[168] Tenopir konstatiert viele der hier erwähnten Probleme im Hinblick auf die Erschließung von Volltexten in einer Zeitschriftendatenbank im Gegensatz zu der Erschließung durch kontrollierte Vokabulare und verweist u.a. auf eine nur halb so große Precision-Rate bei den in den Volltexten indexierten Termen; siehe dazu Tenopir 1985, S. 160. Auch Bertram weist auf das Problem der Irrelevanz vieler der aus Volltexten extrahierten Indexterme hin; vgl. Bertram 2005, S. 43. Des Weiteren merkt sie an, dass die Extraktionsmethode als benennungsorientiertes Verfahren weder Synonymie auflösen noch implizite Inhalte erfassen kann; vgl. ebd., S. 80 f.

kann den Nutzer_innen bei steigenden Dokumentenmengen schon bei der Ausgabe der Treffer einen Hinweis darauf geben, wo ein Thema im Mittelpunkt des Dokuments steht oder wo es nur am Rande und in einer spezifischen Ausprägung eine Rolle spielt.[169] Die Snippets leisten zwar eine solche kontextuelle Einordnung; gleichzeitig steigt jedoch die Zahl an Dokumenten mit gescannten ToC kontinuierlich im Vergleich zu der seit 2016 rückläufigen Zahl intellektuell erschlossener Dokumente. Die Folge hieraus könnte sein, dass gerade die ersten und aktuellsten Treffer, die zu einer Recherche angezeigt werden (sofern die Default-Einstellung der chronologischen Sortierung beibehalten wird), verstärkt durch die Kataloganreicherung aufgefunden werden und damit u.U. in sehr spezifischen Gebrauchskontexten vorkommen, während Dokumente, die als Ganzes intellektuell einem Thema zugewiesen wurden, erst auf den hinteren Seiten der Trefferliste zu finden sind.[170]

[169] Auch Ingwersen/Järvelin weisen auf das Problem hin, dass beim Exact Match keine Gewichtung der Indexterme vorgenommen wird; vgl. Ingwersen/Järvelin 2005, S. 119.

[170] Erste Pre-Tests zu dieser Studie belegen dies. Bei einer ersten Variante der Testanordnung wurde noch ohne den einschränkenden LSW-Suchschlüssel gearbeitet, was zur Folge hatte, dass die ersten Dokumente, die neben den anderen Variablen auch LSW vorhielten, erst nach Durchsicht mehrerer hundert Treffer gefunden wurden.

5. Fazit und Ausblick

In der hier vorgenommenen Studie konnte gezeigt werden, dass lokale Schlagworte einen nicht unerheblichen Anteil am Retrieval haben. Bezogen auf das gesamte Sample aus 80 Suchanfragen mit 697 berücksichtigten Dokumenten waren LSW bei einem Viertel der Dokumente an ihrem Auffinden beteiligt. Und ein Fünftel der Gesamtmenge der Dokumente wäre gar nicht gefunden worden, wenn sie nicht durch LSW intellektuell erschlossen worden wären.

Allerdings bietet sich unter Berücksichtigung der Multilingualität auf Seiten der Suchsprache ein deutlich differenzierteres Bild. So haben die überwiegend deutschsprachigen LSW bei deutschsprachigen Sucheingaben (oder solchen, die als deutschsprachig interpretiert werden können, siehe dazu Kapitel 3.1.2.) eine wesentlich größere Bedeutung und sind für mehr als zwei Fünftel der Treffer verantwortlich. Für deutschsprachige Suchanfragen bedeutet der Wegfall der Schlagworte als einem möglichen Sucheinstieg also erhebliche Einbußen beim Retrieval. Bei fremdsprachigen Sucheingaben hingegen ist ihr Anteil verschwindend gering.
Der Vergleich der Häufigkeitsverteilung der aufgefundenen Dokumente auf die verschiedenen Sucheinstiege bei Einzelbetrachtung beider Samples legt die Vermutung nahe, dass die Sprache, die bei der Suche im mehrsprachigen Bestand des IAI gewählt wird, deutliche Auswirkungen auf den Anteil der Dokumente hat, die durch LSW gefunden werden. Die geringe Menge getesteter Suchanfragen sowie das Ausstehen weitergehender statistischer Analysen erlauben an dieser Stelle jedoch keine abschließenden Rückschlüsse, auch wenn die hier gewonnenen Daten eine sehr eindeutige Tendenz zeigen.
In dieser Arbeit werden ausschließlich Aussagen zu Mengenanteil und Verteilung der Trefferdokumente getroffen; vielfältige andere Aspekte, die grundsätzlich oder im Hinblick auf die Dimension der Multilingualität von Interesse wären, konnten hier nicht weiter berücksichtigt werden. Die aufgrund der manuellen Auswertung erforderliche detaillierte Dokumentation und Aufzeichnung der Daten würde es aber durchaus möglich machen, einigen dieser Aspekte in einer Folgestudie nachzugehen. Die zeitintensive manuelle Dokumentation und Auswertung großer Datenmengen mögen sicherlich nicht das Mittel der Wahl sein und bergen ein nicht unerhebliches Fehlerrisiko; andererseits ermöglichen sie Einsichten, die bei einer automatischen Auswertung verborgen bleiben. Die Anwendung sehr spezifischer Indexierungsparameter bei bestimmten Indexfeldern, die darum keine Treffer erzeugen, sprachliche Ambiguitäten bei den Indextermen aus der Volltextindexierung (z.B. „Barroco“ als Name) oder Korrelationen mit der Objektsprache der aufgefundenen

Dokumente wären ansonsten u.U. nicht in den Fokus gerückt und als Desiderat weiterer Untersuchungen identifiziert worden.

Bezogen auf den Aspekt der Multilingualität wäre v.a. eine Untersuchung der Objektsprache, d.h. der Sprache, in der die Dokumente selbst vorliegen – und damit auch Teile ihrer bibliographischen Metadaten wie etwa der Titel oder die Indexterme aus den ToC –, von großem Interesse, da sie es ermöglichen würde, eine Korrelation zwischen Eingabe-, Indexierungs- und Objektsprache herzustellen. D.h. es könnte nicht nur eruiert werden, welchen Einfluss die *Suchsprache* auf die Beteiligung der verschiedenen Sucheinstiege beim Auffinden der Dokumente hat, sondern auch, ob ein Zusammenhang zwischen Art und sprachlicher Verfasstheit der Indexterme und der *Objektsprache* der Dokumente, die diese Terme enthalten, besteht.

In den hier exemplarisch untersuchten Einzelfällen, die durch ihre geringe Zahl jedoch keine Repräsentativität beanspruchen können, erweisen sich die lokalen Schlagworte, sofern sie sich mit den Suchanfragen der Nutzer_innen decken, als ein geeignetes Instrument, um verschiedensprachige Dokumente thematisch zu bündeln. Die Achillesverse ist hier allerdings, dass die überwiegend auf Deutsch vorliegenden Schlagworte aus dem lokalen Thesaurus es erforderlich machen, dass die multlingualen Nutzer_innen des IAI ihre Suchanfragen in deutscher Sprache stellen. Bei den fremdsprachigen Suchtermen zeigt sich bei den untersuchten Fallbeispielen hingegen, dass Nutzer_innen, die nicht auf Deutsch suchen, von der homogenisierenden Wirkung der Schlagworte kaum profitieren können.

Bei Suchanfragen, die eindeutig einer Fremdsprache zugeordnet werden können, ist ausgehend von den hier vorgestellten Ergebnissen anzunehmen, dass die Trefferdokumente in der Mehrheit durch die Volltextindexierung der ToC aufgefunden werden. Allerdings ermöglicht die rein benennungsorientierte Indexierung durch Termextraktion keinen Zugang zu Dokumenten in anderen Sprachen. Diese können aber durchaus den Informationsbedürfnissen der Nutzer_innen entsprechen: entweder, da diese möglichst vollständig die Forschungsliteratur zu einem Thema erfassen wollen, auch wenn ihnen diese sprachlich unzugänglich ist, oder – im Falle der „polyglots“ – da die Nutzer_innen abseits ihrer präferierten Suchsprache auch andere Sprachen dominieren.

Thesenhaft formuliert könnte man also sagen, dass die lokalen Schlagworte des IAI für deutschsprachige Nutzer_innen Dokumente unabhängig von der Sprache, in der sie verfasst sind, zugänglich machen, während fremdsprachige Nutzer_innen auf die anderen Sucheinstiege angewiesen sind und für eine thematische Homogenisierung ihrer Ergebnisse v.a. auf fremdsprachige Schlagworte aus Fremddaten hoffen müssen.

Ebenfalls nutzbringend wäre eine weitergehende Analyse aller hier dokumentierten Sucheinstiege auf ihren Anteil am Retrieval. Für die Kataloganreicherung und die SW aus

Fremddaten wurde dies im Ansatz getan. Insbesondere die Untersuchung von Korrelationen zwischen den verschiedenen Sucheinstiegen – beispielsweise Erhebungen dazu, in Kombination mit welchem weiteren Sucheinstieg die LSW (oder andere Variablen) bei den Trefferdokumenten am häufigsten indexiert werden – könnte weitere Einblicke in die Bedeutung einzelner Variablen für das Retrieval geben. So ließe sich beispielsweise das Verhältnis von lokalen Schlagworten und Schlagworten aus Fremddaten genauer auf Kongruenz bzw. Komplementarität ausleuchten.

Ausgehend von diesen weitergehenden Überlegungen lassen sich zwei Hypothesen formulieren, denen in einer Folgestudie nachgegangen werden könnte:

- Die Objektsprachen der aufgefundenen Dokumente variieren, je nachdem in welchen Sprachen die Suchanfragen verfasst sind, wobei anzunehmen ist, dass fremdsprachige Suchanfragen keine oder kaum deutschsprachige Dokumente auffinden.
- Die Übernahme von Fremddaten und die Kataloganreicherung durch eingescannte ToC führen zu einer Ausweitung fremdsprachiger Retrievalergebnisse: zum einen, da auch fremdsprachige kontrollierte Vokabulare indexiert werden, und zum anderen, da bei der Volltextindexierung der ToC auch fremdsprachige Terme extrahiert werden.

Wie Wyly in seiner Studie von 1996 betont, ist es wichtig die verschiedenen Sucheinstiege datenbasiert auszuwerten, um ihren Wert für das Retrieval angemessen einschätzen zu können und eine Entscheidung über den Umgang mit ihnen zu treffen.[171] Für den hier betrachteten Sucheinstieg der lokalen Schlagworte kann man dementsprechend folgern, dass sie als eine intellektuelle Form der Inhaltserschließung zwar zeit- und damit kostenintensiv sein mögen; ihr Nutzen erweist sich bezogen auf die Gesamtmenge der untersuchten Anfragen jedoch weiterhin als hoch und insbesondere bei deutschsprachigen Anfragen als beinahe unverzichtbar, da ein sehr großer Teil der Dokumente ohne sie nicht aufgefunden würde. Zwar kann angenommen werden, dass – wie in der Forschungsliteratur konstatiert – die wenigsten Nutzer_innen gezielt mit Schlagworten suchen, etwa über das entsprechend auswählbare Suchfeld; durch ihre Indexierung haben sie aber dennoch Auswirkungen auf die Trefferergebnisse, die zu einer Suchanfrage gefunden werden. Darüber hinaus leisten sie eine Homogenisierung der Ergebnisse und eine sprachliche Disambiguierung, die in der jetzigen Form der Kataloganreicherung nicht geleistet wird. Der Einsatz kontrollierter Vokabulare scheint damit auch trotz Kataloganreicherung sinnvoll und keinesfalls obsolet.

[171] Vgl. Wyly 1996, S. 211 f. Ebenso argumentieren Ingwersen/Järvelin 2005, S. 134.

Neben der rein quantitativen Bedeutung der LSW, zumindest für eines der beiden untersuchten Sprachsamples – und, so lässt sich vermuten, eine von zwei unterschiedlichen Nutzer_innengruppen –, sind folglich auch qualitative Aspekte zu beachten, die hier jedoch nicht weitergehend untersucht werden konnten.
So wäre eine Relevanzbewertung der aufgefundenen Dokumente sicherlich eine gute ergänzende Methode zu der hier vorgenommenen, rein quantitativen Auswertung der Verteilung der Fundstellen der Suchterme über den Index. Eine nutzerorientierte Relevanzbewertung könnte die hier gewonnenen Ergebnisse noch weiter differenzieren, wobei auch die Informationsbedürfnisse der Suchenden genauer ermittelt werden könnten. Auch eine Erweiterung auf Tests mit Mehrwortanfragen, also der Eingabe von mehr als einem Suchterm, könnte dabei helfen, einerseits die Suchanfragen eindeutiger einer Sprache zuzuordnen und andererseits die Informationsbedürfnisse der Suchenden zu konkretisieren, die in Einwortanfragen oft sehr allgemein erscheinen oder vieldeutig interpretierbar sind. Über eine differenziertere Auswertung der aus Logfiles gewonnen Daten hinaus wäre jedoch eine Interaktion mit konkreten Nutzer_innen sicherlich sinnvoll.[172] Durch ein solches Vorgehen könnten die hier vorgenommenen Analysen um eine entscheidende Perspektive ergänzt werden, die einem rein systemorientierten Ansatz notwendigerweise fehlt: die Sicht der Nutzer_innen auf die im Retrieval gewonnenen Ergebnisse.[173]

Die Erkenntnisse aus dieser Studie zeigen jedoch unabhängig von möglichen weiteren Untersuchungen, dass Multilingualität eine große Herausforderung darstellt, insbesondere für Gedächtnisinstitutionen mit einem mehrsprachigen Bestand, der von Nutzer_innen mit verschiedenen Suchsprachen genutzt wird. Der große Unterschied beim Anteil der LSW am Retrieval je nach gewählter Suchsprache legt nahe, dass hier jeweils unterschiedliche Bedürfnislagen vorherrschen und dementsprechend verschiedene Strategien zum Einsatz kommen müssen, um diesen Bedürfnissen gerecht zu werden und die Retrievalergebnisse unabhängig von der genutzten Sprache zu homogenisieren. Am IAI ist aktuell eine multilinguale Indexerweiterung für die Kataloganreicherung geplant. Dies ist sicherlich ein guter und wichtiger Schritt um das – durch die hier gewonnenen Zahlen gestützte – große Potential der Kataloganreicherung beim Auffinden von Dokumenten nachhaltig zu nutzen und die Dokumente auch Nutzer_innen mit anderen Suchsprachen zugänglich zu machen.

[172] Auf die kognitiven und kontextuellen Grenzen der TLA wurde in Kapitel 2.6. genauer eingegangen.

[173] Um hier zu differenzierten Aussagen zu gelangen, wäre insbesondere eine graduelle Relevanzbewertung wünschenswert, d.h. die Einordnung der Relevanz auf einer Skala und nicht bloß eine binäre Bewertung nach relevant/nicht relevant; zu diesem Vorgehen vgl. Ingwersen/Järvelin 2005, S. 129 f., 178.

Wie hier gezeigt werden konnte, ist der Anteil der lokalen Schlagworte beim Auffinden von Dokumenten für die deutschsprachigen Anfragen ebenfalls sehr groß. Fremdsprachige Anfragen profitieren von diesem Potential allerdings kaum und damit auch nicht von all den weiteren Vorzügen kontrollierter Vokabulare. Auch hier wäre eine multilinguale Erweiterung, etwa durch Crosskonkordanzen mit anderen Thesauri oder eine Übersetzung ins Deutsche bereits auf der Ebene der Suchanfragen, sicherlich von großem Nutzen, um die unbestrittenermaßen sehr zeit- und kostenintensive intellektuelle Inhaltserschließung nachzunutzen und ihr Potential auch für eine Verbesserung des Retrievals bei fremdsprachigen Suchen einzusetzen.[174] Das Ziel der babylonischen Bibliothek, wie sie Borges beschreibt, ist es ja gerade, im Labyrinth der Buchstaben und Sprachen durch die ordnende Kraft der Systematik noch jedes kostbare Buch auffindbar zu machen.

[174] Eine Grundvoraussetzung für die sinnvolle Nachnutzung des lokalen Thesaurus des IAI wäre jedoch die kontinuierliche Pflege und Aktualisierung der darin enthaltenen Deskriptoren, insbesondere der nur noch sehr sporadisch veränderten Sachschlagworte, um so auch neuen, aus aktuellen Forschungsfragen erwachsenden Informationsbedürfnissen der Nutzer_innen gerecht werden zu können.

6. Literaturverzeichnis

Bertram, Jutta (2005): Einführung in die inhaltliche Erschließung. Grundlagen, Methoden, Instrumente. Würzburg: Ergon-Verlag (Content and Communication, 2).

Blecic, Deborah D.; Bangalore, Nirmala S.; Dorsch, Josephine L.; Henderson, Cynthia L.; Koenig, Melissa H.; Weller, Ann C. (1998): Using Transaction Log Analysis to Improve OPAC Retrieval Results. In: *College & Research Libraries* 59 (1), S. 39–50.

Boltzendahl, Sabine (2003): Ontologien in digitalen Bibliotheken unter dem Schwerpunkt Inhaltserschließung und Recherche. Berlin: Institut für Bibliothekswissenschaft der Humboldt-Universität zu Berlin (Berliner Handreichungen zur Bibliotheks- und Informationswissenschaft, 111).

Borges, Jorge Luis (2005): La biblioteca de Babel. In: Jorge Luis Borges: Narraciones. 16. Auflage. Hg. von Marcos Ricardo Barnatán. Madrid: Cátedra (Letras hispánicas, 123), S. 105–114.

Borges, Jorge Luis (2007): Die Bibliothek von Babel. In: Jorge Luis Borges: 25. August 1983 und andere Erzählungen. Mit einem Vorwort von Martin Gregor-Dellin. Aus dem Spanischen von Maria Bamberg. Frankfurt am Main [u.a.]: Büchergilde Gutenberg (Die Bibliothek von Babel, 5), S. 19–32.

Borgman, Christine L. (1986): Why Are Online Catalogs Hard to Use? Lessons Learned from Information-Retrieval Studies. In: *Journal of the Association for Information Science* 37 (6), S. 387–400.

Borgman, Christine L. (1996): Why Are Online Catalogs Still Hard to Use? In: *Journal of the American Society for Information Science* 47 (7), S. 493–503.

Bortz, Jürgen; Schuster, Christof (2010): Statistik für Human- und Sozialwissenschaftler. Mit 70 Abbildungen und 163 Tabellen. 7., vollständig überarbeitete und erweiterte Auflage. Berlin/Heidelberg: Springer (Springer-Lehrbuch).

Bosca, Alessio; Dini, Luca (2009): Cacao Project at the Logclef Track. In: Carol Peters und Nicola Ferro (Hg.): Working Notes for CLEF 2009 Workshop. Corfù, Greece, September 30 - October 2, 2009 (CEUR Workshop Proceedings, 1175), o.S. Online verfügbar unter: http://ceur-ws.org/Vol-1175/, zuletzt geprüft am 11.03.2020.

Castro Valle, Jorge; Göbel, Barbara; Lehmann, Klaus-Dieter (2005): 75 Jahre Ibero-Amerikanisches Institut. Berlin: Ibero-Amerikanisches Institut.

Ceynowa, Klaus (2017): In Frankfurt lesen jetzt zuerst Maschinen. In: *Frankfurter Allgemeine*, 31.07.2017, o.S. Online verfügbar unter: https://www.faz.net/-gr0-909kq, zuletzt geprüft am 11.03.2020.

Dawson, Andy; Williams, Pete; Gunter, Barrie (2006): Triangulating Qualitative Research and Computer Transaction Logs in Health Information Studies. In: *Aslib Proceedings* 58 (1/2), S. 129–139.

Flaherty, Patricia (1993): Transaction Logging Systems. A Descriptive Summary. In: *Library Hi Tech 11* (2), S. 67–78.

Fourie, Ina; Bothma, Theo (2007): Information Seeking. An Overview of Web Tracking and the Criteria for Tracking Software. In: *Aslib Proceedings* 59 (3), S. 264–284.

Fühles-Ubach, Simone; Umlauf, Konrad (2013): Quantitative Methoden. In: Konrad Umlauf, Simone Fühles-Ubach und Michael Seadle (Hg.): Handbuch Methoden der Bibliotheks- und Informationswissenschaft. Bibliotheks-, Benutzerforschung, Informationsanalyse. Berlin [u.a.]: De Gruyter Saur, S. 80-95.

Furnas, George W.; Landauer, Thomas K.; Gomez, Louis M.; Dumais, Susan T. (1987): The Vocabulary Problem in Human-System Communication. In: *Communications of the ACM* 30 (11), S. 964–971.

Gäde, Maria; Petras, Vivien; Stiller, Juliane (2010): Which Log for Which Information? Gathering Multilingual Data from Different Log File Types. In: Maristella Agosti, Nicola Ferro, Carol Peters, Maarten de Rijke und Alan Smeaton (Hg.): Multilingual and Multimodal Information Access Evaluation. International Conference of the Cross-Language Evaluation Forum, CLEF 2010. Padua, Italy, September 20-23, 2010. Proceedings. Berlin [u.a.]: Springer (Lecture Notes in Computer Science, 6360), S. 70-81.

Garrett, Jeffrey (2006): KWIC and Dirty? Human Cognition and the Claims of Full-Text Searching. In: *The Journal of Electronic Publishing* 9 (1), o.S. Online verfügbar unter: http://dx.doi.org/10.3998/3336451.0009.106, zuletzt geprüft am 11.03.2020.

Garrett, Jeffrey (2007): Subject Headings in Full-Text Environments. The ECCO Experiment. In: *College & Research Libraries* 68 (1), S. 69–81.

Greifeneder, Elke (2013): Benutzerforschung. In: Konrad Umlauf, Simone Fühles-Ubach und Michael Seadle (Hg.): Handbuch Methoden der Bibliotheks- und Informationswissenschaft. Bibliotheks-, Benutzerforschung, Informationsanalyse. Berlin [u.a.]: De Gruyter Saur, S. 257–283.

Griffiths, Jillian R.; Hartley, Richard J.; Willson, Jonathan P. (2002): An Improved Method of Studying User-System Interaction by Combining Transaction Log Analysis and Protocol Analysis. In: *Information Research* 7 (4), o.S. Online verfügbar unter: http://www.informationr.net/ir/7-4/paper139.html, zuletzt geprüft am 11.03.2020.

Gross, Tina; Taylor, Arlene G. (2005): What Have We Got to Lose? The Effect of Controlled Vocabulary on Keyword Searching Results. In: *College & Research Libraries* 66 (3), S. 212–230.

Gross, Tina; Taylor, Arlene G.; Joudrey, Daniel N. (2015): Still a Lot to Lose. The Role of Controlled Vocabulary in Keyword Searching. In: *Cataloging & Classification Quarterly* 53, S. 1–39.

Hauer, Manfred (2004): Durch Content-Ergänzung, maschinelle Indexierung und modernes Information Retrieval können Recherchen in Bibliothekskatalogen deutlich verbessert werden. In: *ABI-Technik* 24 (4), S. 262–268.

Hauer, Manfred (2013): Zur Bedeutung normierter Terminologien in Zeiten moderner Sprach- und Information-Retrieval-Technologien. In: *ABI-Technik* 33 (1), S. 2–6.

Hinrichs, Imma; Milmeister, Gérard; Schäuble, Peter; Steenweg, Helge (2016): Computerunterstützte Sacherschließung mit dem Digitalen Assistenten (DA-2). In: *O-bib* 3 (4), S. 156–185.

Hunter, Rhonda N. (1991): Successes and Failures of Patrons Searching the Online Catalog at a Large Academic Library. A Transaction Log Analysis. In: *RQ* 30 (3), S. 395–402.

IFLA Working Group on Guidelines for Multilingual Thesauri (2009): Guidelines for Multilingual Thesauri. Den Haag: International Federation of Library Associations and Institutions (IFLA) Headquarters (IFLA Professional Reports, 115).

Ingwersen, Peter; Järvelin, Kalervo (2005): The Turn. Integration of Information Seeking and Retrieval in Context. Dordrecht: Springer (Kluwer International Series on Information Retrieval, 18).

Junger, Ulrike (2015): Quo vadis Inhaltserschließung der Deutschen Nationalbibliothek? Herausforderungen und Perspektiven. In: *O-bib* 2 (1), S. 15–26.

Kaske, Neal K. (1993): Research Methodologies and Transaction Log Analysis. Issues, Questions, and a Proposed Model. In: *Library Hi Tech* 11 (2), S. 79–86.

Kasprzik, Anna (2014): Automatisierte und semiautomatisierte Klassifizierung. Eine Analyse aktueller Projekte. In: *Perspektive Bibliothek* 3 (1), S. 85–110.

Keller, Alice (2015): Einstellung zur (automatischen) Sacherschließung in deutsch- und englischsprachigen Ländern. In: *Bibliotheksdienst* 49 (8), S. 801–813.

Kelly, Diane (2009): Methods for Evaluating Interactive Information Retrieval Systems with Users. Boston, Massachusetts [u.a.]: Now Publishers (Foundations and Trends in Information Retrieval, 3, 1/2).

Kempf, Andreas Oskar (2013): Automatische Inhaltserschließung in der Fachinformation. Eine Evaluation zur maschinellen Indexierung sozialwissenschaftlicher Forschungsliteratur. In: *Information - Wissenschaft & Praxis* 64 (2/3), S. 96–106.

Kempf, Andreas Oskar; Zapilko, Benjamin (2013): Normdatenpflege in Zeiten der Automatisierung. Erstellung und Evaluation automatisch aufgebauter Thesaurus-Crosskonkordanzen. In: *Information - Wissenschaft & Praxis* 64 (4), S. 199–207.

Kurth, Martin (1993): The Limits and Limitations of Transaction Log Analysis. In: *Library Hi Tech* 11 (2), S. 98–104.

Larson, Ray R. (1991): The Decline of Subject Searching. Long-Term Trends and Patterns of Index Use in an Online Catalog. In: *Journal of the American Society for Information Science* 42 (3), S. 197–215.

Lepsky, Klaus (1994): Maschinelles Indexieren zur Verbesserung der sachlichen Suche im OPAC. DFG-Projekt an der Universitäts- und Landesbibliothek Düsseldorf. In: *Bibliotheksdienst* 28 (8), S. 1234–1242.

Lepsky, Klaus; Zimmermann, Harald H. (2000): Katalogerweiterung durch Scanning und automatische Dokumenterschließung. Ergebnisse des DFG-Projekts KASCADE. In: *Zeitschrift für Bibliothekswesen und Bibliographie* 47, S. 305–316.

Loth, Klaus (2004): Thematische Abfrage einer dreisprachigen Datenbank mit computerlinguistischen Komponenten. In: *ABI-Technik* 24 (4), S. 294–300.

Malits, Andrea; Schäuble, Peter (2014): Der Digitale Assistent. Halbautomatisches Verfahren der Sacherschließung in der Zentralbibliothek Zürich. In: *ABI-Technik* 34 (3/4), S. 132–143.

Mann, Thomas (2005): Research at Risk. In: *Library Journal* 130 (12), S. 38-40.

Markey, Karen (1984): Subject Searching in Library Catalogs. Before and After the Introduction of Online Catalogs. Dublin, Ohio: Online Computer Library Center (OCLC Library, Information, and Computer Science Series, 4).

Mayr, Philipp; Petras, Vivien (2008a): Building a Terminology Network for Search. The KoMoHe Project. In: Jane Greenberg und Wolfgang Klas (Hg.): Metadata for Semantic and Social Applications. Proceedings of the International Conference on Dublin Core and Metadata Applications, Berlin, 22-26 September 2008, DC 2008: Berlin, Germany. Göttingen: Universitätsverlag Göttingen, S. 177–182.

Mayr, Philipp; Petras, Vivien (2008b): Cross-Concordances. Terminology Mapping and Its Effectiveness for Information Retrieval. In: *International Cataloguing and Bibliographic Control* 38 (3), S. 43–52.

McKinin, Emma Jean; Sievert, MaryEllen C.; Johnson, E. Diane; Mitchell, Joyce A. (1991): The Medline/Full-Text Research Project. In: *Journal of the American Society for Information Science* 42 (4), S. 297–307.

Nicholas, David; Huntington, Paul; Jamali, Hamid R.; Tenopir, Carol (2006): Finding Information in (Very Large) Digital Libraries. A Deep Log Approach to Determining Differences in Use According to Method of Access. In: *The Journal of Academic Librarianship* 32 (2), S. 119–126.

Nohr, Holger (2001): Automatische Indexierung. Einführung in betriebliche Verfahren, Systeme und Anwendungen. Potsdam: Verlag für Berlin-Brandenburg (Materialien zur Information und Dokumentation, 13).

Nowick, Elaine A.; Mering, Margaret (2003): Comparisons Between Internet Users' Free-Text Queries and Controlled Vocabularies. A Case Study in Water Quality. In: *Technical Services Quarterly* 21 (2), S. 15–32.

Nübel, Rita; Schmidt, Paul (2003): Automatische mehrsprachige Indexierung mit dem AUTINDEX System. In: Ralph Schmidt (Hg.): Competence in Content. 25. Online-Tagung der DGI, Frankfurt am Main, 3. bis 5. Juni 2003. Proceedings. Frankfurt am

Main: Deutsche Gesellschaft für Informationswissenschaft und Informationspraxis (DGI), S. 88–99.

Oard, Douglas W. (2009): Multilingual Information Access. In: Marcia J. Bates und Mary Niles Maack (Hg.): Encyclopedia of Library and Information Sciences. 3. Auflage. Boca Raton, Florida: CRC Press, o.S. Online verfügbar unter: http://terpconnect.umd.edu/~oard/pdf/elis09.pdf, zuletzt geprüft am 11.03.2020.

Oberhauser, Otto; Labner, Josef (2003): OPAC-Erweiterung durch automatische Indexierung. Empirische Untersuchung mit Daten aus dem Österreichischen Verbundkatalog. In: *ABI-Technik* 23 (4), S. 305–314.

Peters, Thomas A. (1989): When Smart People Fail. An Analysis of a Transaction Log of an Online Public Access Catalog. In: *The Journal of Academic Librarianship* 15 (5), S. 267–273.

Peters, Thomas A. (1993): The History and Development of Transaction Log Analysis. In: *Library Hi Tech* 11 (2), S. 41–66.

Peters, Thomas A. (1996): Using Transaction Log Analysis for Library Management Information. In: *Library Administration and Management* 10, S. 20–25.

Petras, Vivien (2013): Methoden für die Evaluation von Informationssystemen. In: Konrad Umlauf, Simone Fühles-Ubach und Michael Seadle (Hg.): Handbuch Methoden der Bibliotheks- und Informationswissenschaft. Bibliotheks-, Benutzerforschung, Informationsanalyse. Berlin [u.a.]: De Gruyter Saur, S. 368-386.

Petras, Vivien; Ferro, Nicola; Gäde, Maria; Isaac, Antoine; Kleineberg, Michael; Masiero, Ivano; Nicchio, Mattia; Stiller, Juliane (2012): Cultural Heritage in CLEF (CHiC) Overview 2012. In: Pamela Forner, Jussi Karlgren und Christa Womser-Hacker (Hg.): Third International Conference of the Cross-Language Evaluation Forum, CLEF. Working Notes for CLEF 2012 Conference. Rome, Italy, September 17-20, 2012 (CEUR Workshop Proceedings, 1178), o.S. Online verfügbar unter: http://ceur-ws.org/Vol-1178/, zuletzt geprüft am 11.03.2020.

Priemer, Burkhard (2004): Logfile-Analysen. Möglichkeiten und Grenzen ihrer Nutzung bei Untersuchungen zur Mensch-Maschine-Interaktion. In: *MedienPädagogik: Zeitschrift für Theorie und Praxis der Medienbildung* 4 (Occasional Papers - Einzelbeiträge), S. 1–23. Online verfügbar unter: https://doi.org/10.21240/mpaed/00/2004.06.02.X, zuletzt geprüft am 11.03.2020.

Rädler, Karl (2004): In Bibliothekskatalogen "googlen". Integration von Inhaltsverzeichnissen, Volltexten und WEB-Ressourcen in Bibliothekskataloge. In: *Bibliotheksdienst* 38 (7/8), S. 927–939.

Recker, Ingrid; Ronthaler, Marc; Zillmann, Hartmut (1996): OSIRIS. Osnabrück Intelligent Research Information System – ein Hyperbase Front End System für OPACs. In: *Bibliotheksdienst* 30 (5), S. 833–848.

Ronthaler, Marc; Zillmann, Hartmut (1998): Literaturrecherche mit OSIRIS. Ein Test der OSIRIS-Retrievalkomponente. In: *Bibliotheksdienst* 32 (7), S. 1203–1212.

Rowley, Jennifer (1994): The Controlled versus Natural Indexing Languages Debate Revisited. A Perspective on Information Retrieval Practice and Research. In: *Journal of Information Science* 20 (2), S. 108–118.

Sandore, Beth (1993): Applying the Results of Transaction Log Analysis. In: *Library Hi Tech* 11 (2), S. 87–97.

Sandore, Beth; Flaherty, Patricia; Kaske, Neal K.; Kurth, Martin; Peters, Thomas (1993): A Manifesto Regarding the Future of Transaction Log Analysis. In: *Library Hi Tech* 11 (2), S. 105–106.

Scheven, Esther; Nadj-Guttandin, Julijana (Hg.) (2017): Regeln für die Schlagwortkatalogisierung. RSWK. 4., vollständig überarbeitete Auflage. Frankfurt am Main: Deutsche Nationalbibliothek. Online verfügbar unter: urn:nbn:de:101-2017011305, zuletzt geprüft am 11.03.2020.

Schlittgen, Rainer (1993): Einführung in die Statistik. Analyse und Modellierung von Daten. 4., überarbeitete und erweiterte Auflage. München/Wien: Oldenbourg.

Schöning-Walter, Christa (2010): PETRUS. Prozessunterstützende Software für die digitale Deutsche Nationalbibliothek. In: *Dialog mit Bibliotheken* 22 (1), S. 15–19.

Schöning-Walter, Christa (2011): Automatische Erschließungsverfahren für Netzpublikationen. Zum Stand der Arbeiten im Projekt PETRUS. In: *Dialog mit Bibliotheken* 23 (1), S. 31–36.

Schulz, Ursula (1994): Was wir über OPAC-Nutzer wissen. Fehlertolerante Suchprozesse in OPACs. In: *ABI-Technik* 14 (4), S. 299–310.

Scott, Jane; Trimble, Jeffrey A.; Fallon, L. Fleming (1995): @*&#@ This Computer and the Horse It Rode in on. Patron Frustration and Failure at the OPAC. In: Richard AmRhein (Hg.): Continuity & Transformation. The Promise of Confluence. Proceedings of the Seventh National Conference of the Association of College and Research Libraries, Pittsburgh, Pennsylvania, March 29 - April 1, 1995. Chicago, Illinois: Association of College and Research Libraries, S. 247–256.

Siegmüller, Renate (2007): Verfahren der automatischen Indexierung in bibliotheksbezogenen Anwendungen. Berlin: Institut für Bibliothekswissenschaft der Humboldt-Universität zu Berlin (Berliner Handreichungen zur Bibliotheks- und Informationswissenschaft, 214).

Sridhar, M. S. (2004): Subject Searching in the OPAC of a Special Library. Problems and Issues. In: *OCLC Systems & Services: International Digital Library Perspectives* 20 (4), S. 183–191.

Stiller, Juliane; Gäde, Maria; Petras, Vivien (2010): Ambiguity of Queries and the Challenges for Query Language Detection. In: Martin Braschler, Donna Harman, Emanuele Pianta und Nicola Ferro (Hg.): CLEF 2010 - Conference on Multilingual and Multimodal Information Access Evaluation. Working Notes for CLEF 2010 Conference. Padua, Italy, September 22-23, 2010 (CEUR Workshop Proceedings, 1176), o.S. Online verfügbar unter: http://ceur-ws.org/Vol-1176/, zuletzt geprüft am 11.03.2020.

Stiller, Juliane; Gäde, Maria; Petras, Vivien (2013): Multilingual Access to Digital Libraries. The Europeana Use Case. In: *Information - Wissenschaft & Praxis* 64 (2/3), S. 86–95.

Stiller, Juliane; Király, Péter (2017): Multilinguality of Metadata. Measuring the Multilingual Degree of Europeana's Metadata. In: Maria Gäde, Violeta Trkulja und Vivien Petras (Hg.): Everything Changes, Everything Stays the Same? Understanding Information Spaces. Proceedings of the 15th International Symposium of Information Science (ISI 2017); Berlin, Germany, 13th-15th March 2017. Glückstadt: Verlag Werner Hülsbusch (Schriften zur Informationswissenschaft, 70), S. 164–176.

Stiller, Juliane; Petras, Vivien; Gäde, Maria; Isaac, Antoine (2014): Automatic Enrichments with Controlled Vocabularies in Europeana. Challenges and Consequences. In: Marinos Ioannides, Nadia Magnenat-Thalmann, Eleanor Fink, Roko Žarnić, Alex-Yianing Yen und Ewald Quak (Hg.): Digital Heritage. Progress in Cultural Heritage. Documentation, Preservation, and Protection. 5th International Conference, EuroMed

2014, Lemessos, Cyprus, November 3-8, 2014; Proceedings. Cham [u.a.]: Springer (Lecture Notes in Computer Science, 8740), S. 238–247.

Taylor, Arlene G. (1995): On the Subject of Subjects. In: *The Journal of Academic Librarianship* 21 (6), S. 484–491.

Tenopir, Carol (1985): Full Text Database Retrieval Performance. In: *Online Review* 9 (2), S. 149–164.

Tillotson, Joy (1995): Is Keyword Searching the Answer? In: *College & Research Libraries* 56 (3), S. 199–206.

Uhlmann, Sandro (2013): Automatische Beschlagwortung von deutschsprachigen Netzpublikationen mit dem Vokabular der Gemeinsamen Normdatei (GND). In: *Dialog mit Bibliotheken* 25 (2), S. 26–36.

Voorbij, Henk J. (1998): Title Keywords and Subject Descriptors. A Comparison of Subject Search Entries of Books in the Humanities and Social Sciences. In: *Journal of Documentation* 54 (4), S. 466–476.

Wiesenmüller, Heidrun; Hinrichs, Imma (2017): Computerunterstützte Inhaltserschließung. Bericht über einen Workshop an der UB Stuttgart – mit einem Exkurs zum neuen Inhaltserschließungskonzept der DNB. In: *O-bib* 4 (3), S. 94–105.

Wyly, Brendan J. (1996): From Access Points to Materials. A Transaction Log Analysis of Access Point Value for Online Catalog Users. In: *Library Resources & Technical Services* 40 (3), S. 211–236.

Zink, Steven D. (1991): Monitoring User Search Success through Transaction Log Analysis. The Wolfpac Example. In: *Reference Services Review* 19 (1), S. 49–56.

7. Anhang

Alle hier aufgeführten Anhänge befinden sich als Forschungsdokumentation auf dem edoc-Server der Humboldt-Universität zu Berlin, mit Ausnahme von Anhang 4, der der Vollständigkeit halber in die Übersicht mit aufgenommen wurde. Sie können unter folgender DOI aufgerufen werden: https://doi.org/doi:10.18452/21113.

Anhang 1: Lokaler Thesaurus IAI

Dieser Anhang besteht aus einem Ordner, in dem der lokale Thesaurus, der vom IAI benutzt wird, enthalten ist. Im Ordner befinden sich PDF-Dokumente mit all den im Thesaurus enthaltenen Deskriptoren, die in 7 Gruppen unterteilt werden: 1) Personenschlagworte, 2) Körperschaftsschlagworte, 3) Titelschlagworte, 4) Sachschlagworte, 5) Geographika, 6) Zeitschlagworte, 7) Formalschlagworte. Die Daten der im CBS gespeicherten Schlagwortnormsätze wurden am 07. und 08.12.2017 aus dem CBS heruntergeladen und in Excel importiert. In den Datensätzen vorhandene Fehler, etwa die falsche Zuordnung eines Schlagworts zu einer der Gruppen, Tippfehler o.Ä., wurden nicht bereinigt.

Anhang 2: Logfile nach der Datenbereinigung

Dieses Excel-Dokument enthält das Logfile, das als Grundlage für die Gewinnung der Purpose Samples diente. Es wurde nach den in Kapitel 3.1.1. beschriebenen Kriterien bereinigt und umfasst 6433 Einträge.

Anhang 3: Indexierungsparameter VZG

Dieses PDF enthält die von der VZG bei der Indexierung angewandten Parameter. Für das LBS des IAI ist die Spalte „PICA +“ relevant, in der die den Katalogfeldern zugeordneten Kategorien angegeben werden. Der Spalte „Schlüssel“ kann entnommen werden, mit welchem Suchschlüssel die verschiedenen Kategorien belegt werden, wobei unter Umständen weitere Bedingungen hinzukommen können, etwa die Materialart.

Anhang 4: Dictionary OCLC

Dieses PDF enthält eine Datei aus dem Dictionary-Verzeichnis von OCLC, die regelt, wie bei der Indexierung mit Sonderzeichen und nicht-lateinischen Buchstaben umgegangen wird. Das Dictionary-Verzeichnis wurde als Anhang zu der Masterarbeit eingereicht, aus der diese Publikation entstanden ist. Das Dokument ist allerdings nicht für eine Veröffentlichung auf

dem edoc-Server der Humboldt-Universität zu Berlin freigegeben und kann daher nicht gemeinsam mit den anderen Anhängen öffentlich zur Verfügung gestellt werden.

Anhang 5: Dokumentation der Testläufe

Dieser Ordner enthält die Dokumentation zu den in dieser Arbeit durchgeführten Tests am OPAC des IAI. Er umfasst 3 Unterordner:

1) Dokumentation Nulltreffer: Darin enthalten sind Screenshots der OPAC-Anzeige nach Eingabe eines Suchterms ohne Trefferergebnisse sowie ein PDF, das Zeitpunkt der Durchführung der Suchen sowie etwaige Besonderheiten dokumentiert.
2) Dokumentation Sample deutschsprachig (siehe Kapitel 3.2.2.).
3) Dokumentation Sample fremdsprachig (siehe Kapitel 3.2.2.).

Anhang 6: Testläufe Sample deutschsprachig ohne Indexierungsparameter

Diese Excel-Tabelle dokumentiert die Auswertung der Indexate der ersten 10 berücksichtigten Dokumente, die zu den in den Testläufen eingegebenen deutschsprachigen Suchanfragen gefunden wurden, wobei hier noch nicht die von der VZG zu Grunde gelegten Indexierungsparameter und das Exact Match berücksichtigt wurden.

In der linken Spalte wurden die getesteten Suchterme mit den zu ihnen aufgefundenen und berücksichtigten Dokumenten bis zu einem Cutoff-Wert von 10 notiert. Die Dokumente werden in der Regel über ihre Signatur identifizierbar gemacht – ist diese nicht vorhanden wird die PPN angegeben. Die in eckige Klammern gefasste Zahl hinter Signatur oder PPN gibt an, welche Position das Dokument in der im OPAC angezeigten Trefferliste einnimmt, sodass eine schnelle Zuordnung zu den in der Dokumentation gespeicherten Screenshots möglich ist.

In den weiteren Spalten werden Angaben zu den 5 hier untersuchten Variablen gemacht. In der ersten Spalte zu jeder Variablen wird durch die Eintragung der Ziffer 1 (grün hinterlegt) markiert, wenn der Suchterm durch diese Variable aufgefunden wurde. Die zweite Spalte jeder Variablen enthält verschiedene Kommentare, z.B. zur Art der Kataloganreicherung; zum Schlagwort, das für das Auffinden verantwortlich war; zu abweichenden Schreibweisen durch Diakritka oder zum LBS-Feld, das indexiert wurde. Vereinzelt wurden auch weitere Besonderheiten eingetragen oder Fälle, in denen zwar ein thematisch passendes Schlagwort vergeben wurde, dieses jedoch nicht mit dem Suchterm übereinstimmte. Diese Angaben dienen teilweise der Möglichkeit, weitergehende Analysen für den internen Gebrauch anschließen zu können, und sind für die hier vorgenommenen Auswertungen häufig nicht relevant.

Bei den beiden Variablen Kataloganreicherung und SW aus Fremddaten kommt außerdem eine dritte Spalte hinzu („Nicht vorhanden“), in der vermerkt wurde, wenn ein Katalogisat nicht über die entsprechende Variable verfügt. Hier wird ebenfalls die Ziffer 1 (rot unterlegt) eingetragen.

Anhang 7: Testläufe Sample fremdsprachig ohne Indexierungsparameter

Für das fremdsprachige Sample wurde analog wie in Anhang 6 vorgegangen.

Anhang 8: Testläufe Sample deutschsprachig

Hier wurde analog wie in Anhang 6 vorgegangen. Allerdings wurden nun bei der Auswertung die Indexierungsparameter der VZG berücksichtigt (siehe dazu Kapitel 3.2.2.). Felder, die diesen Parametern nicht entsprechen, wurden rot hinterlegt, und die Eintragung der Ziffer 1 (grün hinterlegt) wurde entfernt, wenn bei einer der Variablen ausschließlich ein nicht mit „ALL“ indexiertes Feld für den Treffer verantwortlich war. In Fällen, in denen ein Suchterm bei einer Variablen in mehreren Feldern indexiert wurde, sowie bei den Indextermen aus der Kataloganreicherung, die häufig verschiedenen Quellen entstammen, wurde die Markierung der nicht mit „ALL“ indexierten Felder durch rote Schrift nicht systematisch vorgenommen. Es wurden jedoch alle Felder nach bestem Wissen und Gewissen geprüft. D.h. es kann davon ausgegangen werden, dass, sofern die Ziffer 1 (grün hinterlegt) eingetragen wurde, mindestens eines der angegebenen Felder mit „ALL“ indexiert wurde.

Anhang 9: Testläufe Sample fremdsprachig

Für das fremdsprachige Sample wurde analog wie in Anhang 8 vorgegangen.

Anhang 10: Normsätze lokal und aus Fremddaten Sample deutschsprachig

Diese Excel-Tabelle führt zu jedem der getesteten deutschsprachigen Suchterme auf, ob dieser in Normsätzen der LSW oder der SW aus Fremddaten gefunden wurde. Dafür wurde in der entsprechenden Spalte die Ziffer 1 (grün hinterlegt) eingetragen. Im Kommentarfeld der Spalte rechts davon wird das Schlagwort verzeichnet, das für das Auffinden verantwortlich war. Ebenfalls eingetragen werden Funde der Suchterme in Normsätzen aus der Formalerschließung, die in dieser Untersuchung den weiteren bibliographischen Daten zugeordnet werden, da sie nicht in den Bereich der Sacherschließung fallen. Diese Felder wurden darum rot hinterlegt. Die Eintragung der Ziffer „0“ (rot hinterlegt) zeigt an, wenn der

Suchterm im entsprechenden Normsatz zwar aufgefunden wurde, aufgrund der Indexierungsparameter jedoch nicht mit „ALL“ indexiert wurde.

Anhang 11: Normsätze lokal und aus Fremddaten Sample fremdsprachig

Für das fremdsprachige Sample wurde analog wie in Anhang 10 vorgegangen.

Anhang 12: Studentisierung Sample gesamt

Dieses PDF-Dokument zeigt eine Tabelle, in der die für jede einzelne der 80 gestellten Suchanfragen relevanten Daten ausgewertet wurden. Der Grad der Beteiligung der LSW beim Auffinden der Dokumente (ausschließlich durch LSW, durch LSW und weitere Variablen, nicht durch LSW) wurde außerdem in Form studentisierter Werte angegeben (siehe dazu Kapitel 3.2.3.).

Anhang 13: Studentisierung Sample deutschsprachig

Für die 40 deutschsprachigen Suchanfragen wurde analog wie in Anhang 12 vorgegangen.

Anhang 14: Studentisierung Sample fremdsprachig

Für die 40 fremdsprachigen Suchanfragen wurde analog wie in Anhang 12 vorgegangen.

Anhang 15: Auswertung ohne Berücksichtigung der Indexierungsparameter

In diesem PDF-Dokument werden ausgehend von Anhang 6 und 7 die Testläufe aller drei Samples ausgewertet, d.h. ohne Berücksichtigung der Indexierungsparameter. Die Tabellen geben zum einen den Grad der Beteiligung der LSW beim Auffinden der Dokumente an und zum anderen die Verteilung der ausschließlich durch eine einzige Variable aufgefundenen Dokumente über alle 5 Sucheinstiege.